# Population Genetics of Mutation Load and Quantitative Traits in Humans

Yuval B. Simons

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

# ABSTRACT

## Population Genetics of Mutation Load and Quantitative Traits in Humans

Yuval B. Simons

The past fifteen years have seen a revolution in human population genetics. We have gone from anecdotal genetic data from a few individuals at a few genetic loci to an avalanche of genome-wide sequencing data, from many individuals in many different human populations. These new data have opened up many new directions of research in human population genetics. In this work, I explore two such directions.

Genomic data have uncovered that recent changes in human population size have had dramatic effects of on the genomes of different human populations. These effects have raised the question of whether historic changes in population size have led to differences in the burden of deleterious mutations, or mutation load, between different human populations. In Chapter 1 of this thesis, I show that despite earlier arguments to the contrary only minor differences in load are expected and indeed observed between Africans and Europeans.

Over the past decade, genome-wide association studies (GWAS) have begun to systematically identify the genetic variants underlying heritable variation in quantitative traits. The number, frequencies and effect sizes of these variants reflect the selection, and other evolutionary processes, acting on traits. In Chapter 2, I develop a model for traits under pleiotropic, stabilizing selection, relate the model's predictions to GWAS findings, and show that GWAS findings for height and BMI indeed follow model predictions. In Chapter 3, I develop a method to infer the distribution of selection coefficients acting on genome-wide significant associations made by GWAS.

# Table of Contents

# List of Figures

# Acknowledgements

This is a work of perseverance. Over six years have passed from the start of this work in Jerusalem to its completion in California. That period of time has brought with it many changes and challenges, both scientific and personal. I am wholeheartedly grateful for every person who helped (perhaps inadvertently) along the way. I will try to thank everyone but am bound to fail spectacularly (and that's the best way to fail).

First of all, I want to thank my family. I thank my parents and sisters, who did their best to understand what I do. I thank my Uncle Adrian for his enduring support, without which I wouldn't have pulled through. I thank my mother in-law, Michal, for reading every word of every paper I publish. I thank the Simkhay clan – Havi, Jack, Guy, Liat, Evan, Roy, Lindsay and the kids – for making me feel at home. I thank the Donis – Eshed, Sharon, Or, Eran, Itay and Maya – for always being there. Special thanks go to Guy for all of the shows and Emily for all of the music - without you life would be so boring.

I also want to thank all of my friends who provided so much love and support. Tal for all the (long) talks, Divya for helping me to survive the core, Tali and Bob for introducing me to Brooklyn, Youli for all the (mis)adventures, Amy for widening my horizons, Susan for widening my culinary horizons, David and Amir for the politics, Tami and Nadav for coming time and time again to NJ, Maria for the honey, Eduardo for the (barely functioning) car, and Einav for introducing me to ramen(!). I would also like to thank everyone I forgot for not holding a grudge.

Of course, I am greatly indebted to all the people who made the science happen. I want to thank all members of the Sella and Przeworski labs – Eyal, Arbel, Tal, Guy A, Laura, David, Jeremy,

Amy, Amir, Chen Chen, Ziyue, Yuki, Ellen, Priya, Laure, Alva, Ipsita, Zach, Zach, Carla, Claude, Molly S, Felix and Hakhamanesh. Special thanks go to the Jeremy and Laura, the brave members of the Super Barton Journal Club, and Guy Amster, for the technical support. I also thank all the scientists who helped guide me, especially Molly Przeworski, Jonathan Pritchard, Joe Pickrell, Itsik Pe'er, Dick Durbin and Nick Barton.

Last but not least, my advisor Guy Sella, who somehow managed to work with me for over six years.

Above all, I want to thank my wife and daughter:

Abigail, when you're old enough to read this know that though you did your best to prevent the completion of this thesis it wouldn't be worth completing if it wasn't for you.

Ayelet, this is truly all yours. שָׁלִי וְשֶׁלָּכֶם – שֶׁלָּהּ הוּא

# Introduction
## Population genetics of mutation load and quantitative traits in humans

The past fifteen years have seen a revolution in human population genetics [1]. We have gone from anecdotal genetic data from a few individuals at a few genetic loci to an avalanche of genome-wide sequencing data, from many individuals, different human populations, and closely related species. These new data have allowed for many central questions in human population genetics to be systematically addressed. For the first time, recombination rates [2], mutation rates [3] and genetic diversity levels [4] could be methodically estimated along the genome in different human populations. These estimates have led to the discovery of recombination hotspots, to a two-fold reduction in estimates of human mutation rates, and to increasingly detailed inferences about the demographic history of human populations.

The demographic history of human populations has dramatically impacted their genomes. Modern human populations originated in Africa, and most non-African populations appear to trace back to a single migration event, known as the Out-of-Africa exodus [5]. The expansion of humans from Africa to the entire globe is thought to have been accompanied by a series of temporary population size reductions, known as bottlenecks, as evidenced by the decline in human diversity levels with distance from Africa (Figure I1A-B). In addition, most human populations have also experienced a sustained period of explosive population growth over the past 10,000 years (Figure I1C) leaving a pronounced footprint on extant genetic variation [6], as reflected in the abundance of rare variants observed in many human populations (Figure I1D).

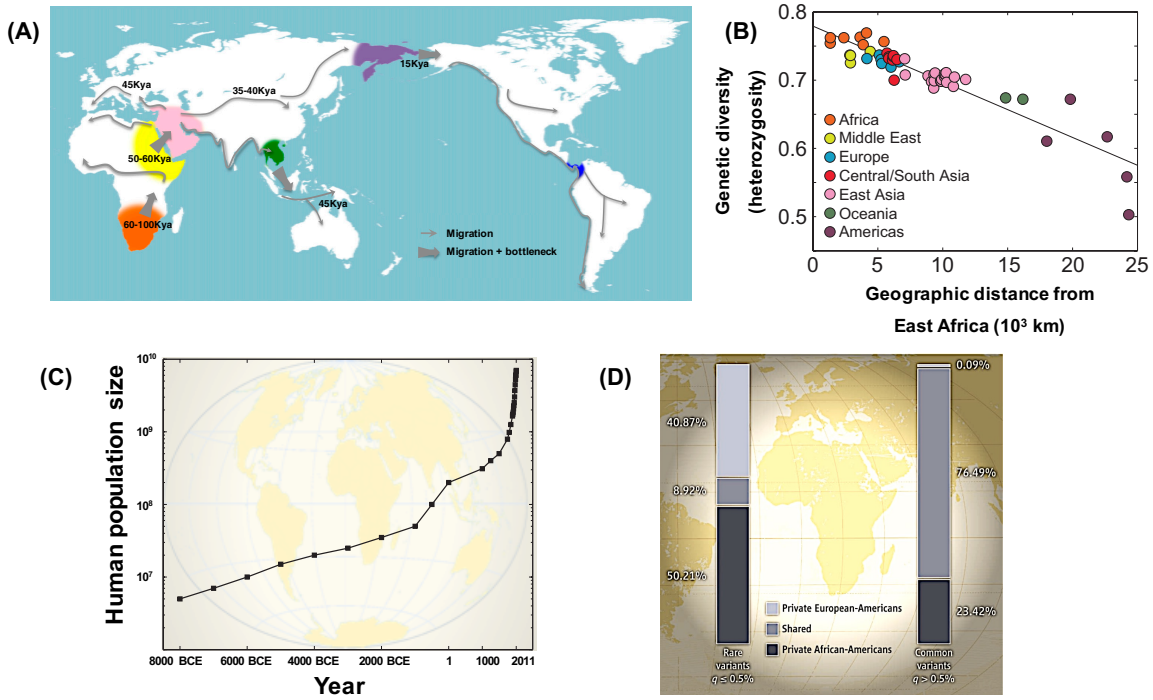



**Figure I1. The genetic impact of recent demographic history on human populations**. (A) World map illustrating the patterns of migrations and bottlenecks during the human expansion out of Africa. Dates given in kilo years ago (Kya). Adapted from Henn et al. [5]. (B) The decline in levels of genetic diversity (measured here by mean heterozygosity) by distance from East Africa. Adapted from DiGiorgio et al. [7]. (C) The estimated census size of global human population by year. Adapted from Keinan and Clark [8]. (D) The abundance of rare private variation in exome data from 1,351 European-Americans and 1,088 African-Americans [9]. Adapted from Casals and Bertranpetit [10].

Since historical changes in population size have created pronounced differences in the abundance and frequencies of genetic variation among extant human populations, they might have also generated differences in the burden of deleterious genetic variation among them. The burden of deleterious mutations is usually quantified in terms of the deleterious mutation load, defined as the average reduction in fitness due to the accumulation of deleterious mutations along the genome. Several studies have argued that selection is less effective during bottlenecks allowing

deleterious alleles to drift to higher frequencies thereby increasing mutation load in bottlenecked populations [11]. Others have argued that difference in load can arise due to the influx of rare variation during population growth [8].

In Chapter 1 of this thesis, done in collaboration with Prof. J.K.Pritchard's group and originally published in Nature Genetics[12], I show that these arguments are largely incorrect, and specifically that only minor differences in load are expected and indeed observed between Africans and Europeans. I use population genetic models and simulations to examine the effects of the Out-of-Africa bottleneck and the recent population growth on the mutation load. I classify the strength of selection acting on deleterious variants into three regimes: strong, weak, and effectively neutral. For variants under strong selection, mutation-selection balance keeps the mutation load constant even under population size changes. The contribution of effectively neutral variants to load is dominated by variants that have been fixed in the population long ago, and the load is therefore insensitive to recent demographic changes. Variants under weak selection might have contributed to a minor difference in load between Africans and Europeans, but because of the short timescales involved this contribution should have been minimal, although weakly deleterious recessive variants may have had a somewhat larger and possibly detectable contribution (see also the recent review paper I coauthored[13]).

I also tested these theoretical predictions using data from two exome (protein coding part of the genome) sequence databases. Since load cannot be measured directly, I used the average number of derived (different than the human-chimp ancestor) protein-altering variants carried by an individual in a given population as a proxy for mutational load. Such derived variants are most often deleterious and therefore mutation load should increase with the number of such variants per individual. Despite massive amounts of data, I find no significant difference in the average

number of derived protein-altering variants carried by individuals of African and European descent, using two different datasets. I also find no signal when I classify the protein-altering variants by their estimated degree of damage (based on a variety of existing methods). Follow up studies have shown these results to hold for a wide variety of extant human populations. In turn, also consistent with our theoretical predictions, some studies point to differences in load between modern humans and extinct archaic human lineages (i.e., Neanderthals and Denisovans), owing to more severe and prolonged reductions in population sizes in these lineages (Figure I2).

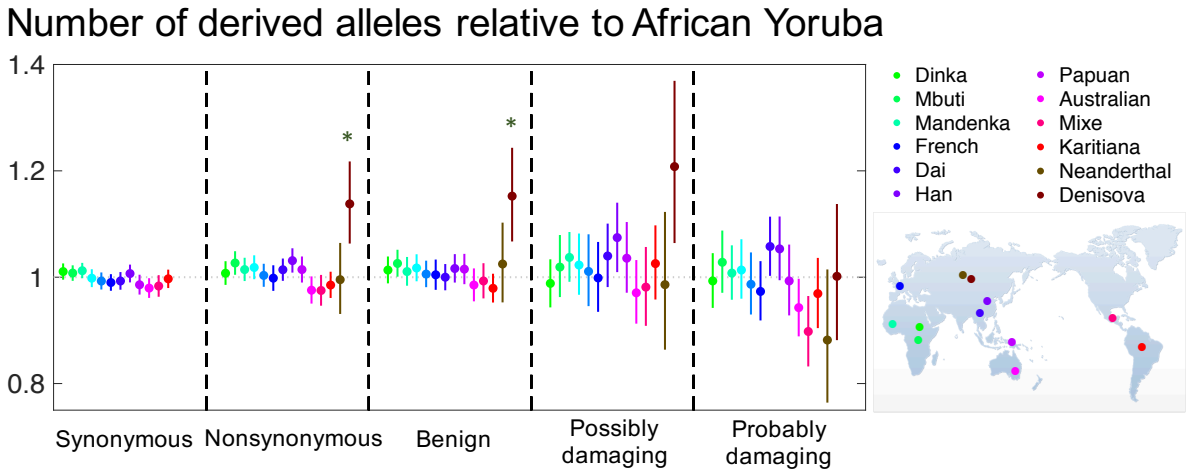



**Figure I2. No significant difference between human populations in the mean number of derived protein-altering variants per individual.** Each population sample was compared with an African Yoruba sample, using data from Do et al. [14]. The grey and brown points correspond to the archaic Neanderthals and Denisovans, respectively. Only the Denisovans are significantly different from Yoruba, as marked by asterisks, although other lines of evidence suggest differences in Neanderthals as well (see ref). The figure is adapted from a review I coauthored on the effects of demography on mutation load in human populations [13], which is not included in this thesis.

The next chapters of my thesis address the evolutionary processes underlying heritable variation in human quantitative traits. Quantitative traits, like height and body mass index (BMI), are continuous in value, generally normally distributed in the population, and are usually highly heritable [15]. The heritable variation in quantitative traits is due to the combined small effects of

many genetic variants -- that is, the traits are polygenic. However, only over the past decade it became possible to systematically identify the genetic variants underlying heritable variation in quantitative traits (Figure I3). In humans, the main method to uncover these variants is genome-wide association studies (GWAS), which have identified many thousands of associations between genetic variants and human quantitative traits [16]. In GWAS, the phenotype (observable trait) of many individuals is measured and their genome is sequenced. The phenotype is then regressed against each accessible site in the genome, in other words the association between each site and the measured trait is estimated (Figure I3A). When the association is strong enough a site is called genome-wide significant (or "GWAS hit") and is considered associated with the trait (Figure I3B).

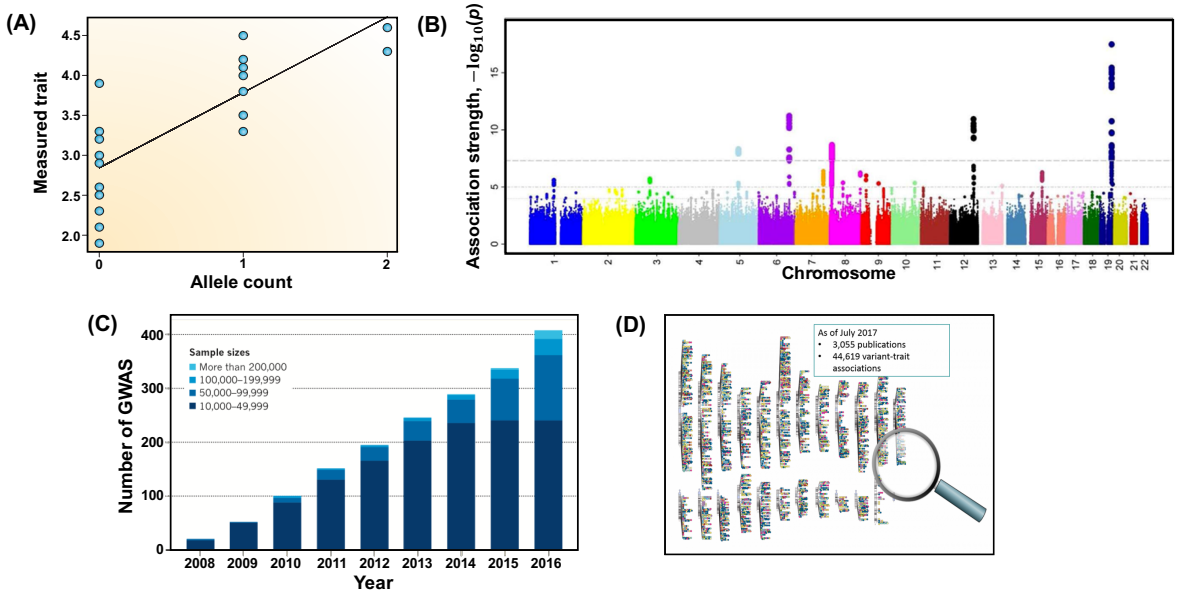



**Figure I3. The advent of genome wide association studies (GWAS).** (A) To identify variants that are associated with a trait, GWAS regress the measured trait vs. allele count in study participants. Taken from Balding [20]. (B) A Manhattan plot, showing the strength of association between each variant in the genome and a trait. Some regions pass a threshold for significance, marked by a horizontal line, and their most associated variants are considered "GWAS hits". Taken from [21]. (C) The increase in number and sample size of GWAS. Taken from Manolio [22]. (D) The genomic position of all catalogued "GWAS hits" as of July 2017. Each vertical bar represents a chromosome, each circle if one GWAS hit, where the colors correspond to different traits [23].

The findings emerging from GWAS are transforming our understanding of heritable variation in quantitative traits. In particular, we have learned a lot about the number variants underlying heritable variation, their distribution in the genome, and their distribution of frequencies and effects sizes, collectively referred to as genetic architecture. GWAS findings suggest that heritable variation arises from numerous variants, most of which have minute contributions to genetic variation and therefore do not reach statistical significance even with study sizes in the hundreds of thousands. Moreover, statistical analysis of GWAS data suggest that these variants are fairly uniformly distributed, common in the population and affect protein regulation rather than disrupt the proteins themselves. GWAS also indicate that traits differ the number, frequency, effect size and distribution of variants underlying trait variation, i.e. in their genetic architecture.

For example, by associating the height of a quarter of a million individuals with their genotype, a recent GWAS has made nearly 700 associations between variants in the genome and human height[17]. These mostly common variants explain roughly 20% of heritable variation in height, suggesting that many more variants are left to be discovered. Indeed, using data for this study it has been estimated that ~4% of all common variants in the genome affect height and that most 100,000 base pair windows in the genome contain at least one common height-affecting variant[18]. While this study size has been sufficient to uncover the genetic basis of a large proportion of heritable variation in height, studies of comparable sizes for other traits, such as BMI[19], have been much less successful and have discovered variants that explain only a small fraction of heritable variation.

An important aspect of interpreting GWAS findings is to understand the evolutionary forces shaping the genetic architecture of traits [24]. In each generation, mutations introduce new variants

that affect quantitative traits into the population. The frequencies of these variants are determined by the selection acting on the trait and by genetic drift. However, a given variant may affect more than one trait, a phenomenon called pleiotropy, and these additional effects can also exert selective pressure on the variant. Therefore, the genetic architectures, as observed in GWAS, reflect the outcomes of these evolutionary processes.

In Chapter 2, originally published in PLOS Biology[25], I develop a mathematical model to describe, from first principles, how these evolutionary forces shape the genetic architecture, creating quantitative testable predictions for GWAS findings. The model assumes that each new variant introduced by mutation affects multiple traits and that these traits are under stabilizing selection, i.e., selection against extreme trait values. When the number of affected traits is sufficiently large (e.g., >10), this model leads to two simple and robust equations. The first describes the expected change in variant frequency each generation as a function of the strength of selection acting on the variant, parametrized by a selection coefficient. The second describes the distribution of effect sizes on a trait of variants with a given selection coefficient. Taken together, these equations imply that when selection is strong the distribution of variant contributions to variance in any given trait follow a universal shape, identical in all traits up to a single scaling constant. Since there is strong evidence that variants indeed affect many traits[26] and since GWAS would currently be underpowered to detect variants under weak selection, we expect the variants discovered by GWAS to follow our predicted universal distribution. However, a GWAS with a given sample size is only powered to detect variants whose contribution to trait variance is above a threshold that is inversely proportional to the sample size. Therefore, GWAS hits will represent the tail of our predicted universal distribution of variant contributions to trait variance.

This prediction can easily be tested against GWAS findings and indeed holds for both height and BMI [17,19], see Figure I4. We fit the distribution of the contributions to trait variance of GWAS hits to our theoretical prediction by estimating a single scaling constant. This scaling constant measures the expected contribution to variance from a single strongly-selected site in the genome. For both traits, I find that the distribution of variant contributions to trait variance fits very well the predicted universal distribution (Figure I4A). The fact that the theory fits the data well allows us to rely on it to make inferences about several important quantities. We can estimate that a strongly selected site makes, on average, comparable contributions to height and BMI but while a large proportion of heritable variation in height (~50%) stems from such sites they account for only a small proportion (~15%) of heritable variation in BMI. This explains why the current GWAS for height captures a much larger proportion of heritable variation than a similarly powered GWAS for BMI[17,19]. In turn, these inferences allow us to predict how many strongly selected variants will be discovered in future GWAS for height and BMI and how much of the heritable variation they will account for (Figure I4B-C).

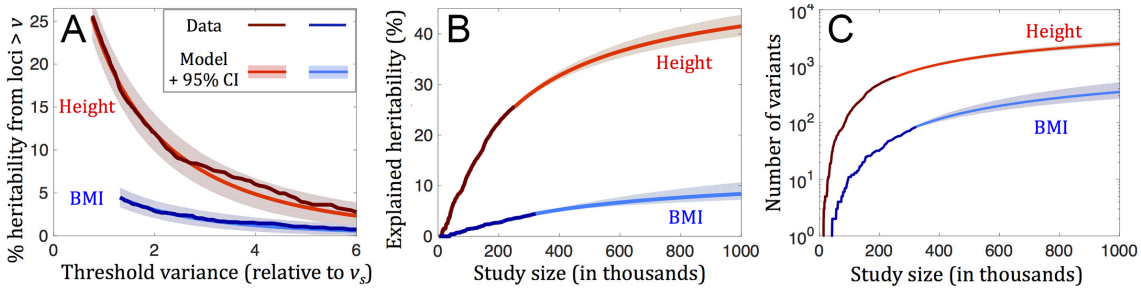



**Figure I4. Our theoretical predictions fit GWAS results for height and BMI[17,19] and allow us to make predictions about future GWAS.** (A) The distribution of the contributions to variance from GWAS hits for height and BMI follows our theoretical prediction. The y-axis represents the heritable variation explained by variants whose contribution to trait variance is above the threshold represented by the x-axis. Prediction of the increase in explained heritability (B) and number of GWAS hits (C) from strongly selected sites as a function of GWAS study size. Taken from Figure 2.5.

We were surprised by how well our theoretical predictions fit GWAS findings in European populations, because our predictions were predicated on a constant population size whereas Europeans have undergone both a severe bottleneck during the Out-of-Africa exodus and recent explosive population growth. As noted in Chapter 1, such changes are known to have affected the distribution of variant frequencies in European populations, including variants that affect height and BMI. These changes in variant frequencies change variant contribution to trait variance and therefore we would expect the distribution of variant contributions to traits, like height and BMI, to deviate from our predictions.

We explain the lack of discrepancy between theory and GWAS findings by the narrow range of selection coefficients discovered in GWAS. Accounting for recent changes in population size, I show, in Chapter 2, that GWAS should only be well-powered to discover variants with a limited range of selection coefficients. Within this range of selection coefficients, the variants contributing most to trait variation are the ones that would have been kept at low frequency if it were not for the bottleneck, but have drifted to high frequencies during the Out-of-Africa bottleneck. The distribution of these variants' contributions to trait variation is well approximated by a constant population size model, reflecting the population size during the Out-of-Africa bottleneck. While these arguments explain why don't see any discrepancy between our constant population size predictions and GWAS findings, they highlight the need for a more quantitative investigation of the effects of demographic changes on trait architecture.

In Chapter 3, I show that by accounting for demographic history, we can use the distribution of variant frequencies and effect sizes to infer the distribution of selection coefficients acting on quantitative trait variation. I use the theoretical framework introduced in Chapter 2 to build a method to infer the distribution of selection coefficients acting on the trait-affecting variants

discovered in GWAS from the variants' frequencies and effect sizes. Allele frequencies provide an upper bound on selection coefficients while effect sizes provide a lower bound, allowing this method to have strong statistical power. Indeed, using extensive simulations, I show that the method is well powered to infer the distribution of selection coefficients given the kind and amount of data that is currently available from human GWAS.

GWAS are only well-powered to discover variants from a limited range of selection coefficients. GWAS hits come from this range of selection coefficients and, since my inference method estimates the distribution of selection coefficients at GWAS hits, my method allows me to estimate this observable range of selection coefficients. Within this range, the inferred distribution of selection coefficients affecting GWAS hits can be converted to an estimate of the number of new trait-affecting mutations arising from each selection coefficient each generation. Such estimates can be used to more accurately predict the number of variants that will be discovered in future GWAS with larger sample sizes, and the proportion of heritable variance that they will account for. This Chapter represents work in progress.

In particular, I am now working on applying this method to GWAS data, with the aim of producing the first inferences of the distribution of selection effects of variants contributing to variation in quantitative traits. To that end, I am using GWAS data from the UK Biobank, which includes genotypes and hundreds of phenotypes from over half a million British individuals[27]. This incredible dataset should allow me to apply my method to many different traits, including morphological traits, such as height and BMI, life history traits, such as age at menarche, and biomedical traits, such as blood lipid levels. Inferring the selection effects of variants affecting many different kinds of traits would allow me to understand the evolutionary origins of differences in the genetic architecture of traits.

# Chapter 1

# The deleterious mutation load is insensitive to recent population history

(Published under Simons et al. Nature Genetics 2014)

## Abstract

**Human populations have undergone dramatic changes in population size in the past 100,000 years, including recent rapid growth. How these demographic events have affected the burden of deleterious mutations in individuals and the frequencies of disease mutations in populations remains unclear. We use population genetic models to show that recent human demography has likely had little impact on the average burden of deleterious mutations. This prediction is supported by two exome sequence datasets showing that individuals of west African and European ancestry carry very similar burdens of damaging mutations. We further show that for many diseases, rare alleles are unlikely to contribute a large fraction of the heritable variation, and therefore the impact of recent growth is likely to be modest. However, for those diseases that have a direct impact on fitness, strongly deleterious rare mutations likely do play an important role, and recent growth will have increased their impact.**

## Introduction

Recent work has highlighted the impact of demographic history on the distribution of human genetic variation. Deep sequencing studies have identified huge numbers of very rare variants in human populations, the consequence of explosive population growth in the past five thousand years[1–6]. Additionally, Europeans and east Asians have a greater fraction of high-frequency variants compared to Africans, likely due to an ancient bottleneck of non-African populations[5, 7, 8–10].

Given these observations, it is natural to ask whether recent demographic history has impacted the burden of genetic disease in modern human populations[3, 6, 11,12]. Keinan and Clark[3] recently hypothesized that "Some degree of genetic risk for complex disease may be due to this recent rapid increase in the number of rare variants in the human population". A second important question concerns the relative importance of rare and common variants in causing disease[13–15]. If much of the genetic variation underlying disease is due to rare variants, then this could help to explain the so-called "missing heritability" of complex traits, and imply that mapping approaches based on deep sequencing will be essential for the dissection of complex traits[16].

## Results

To address these questions, we analyzed a theoretical model with a large number of bi-allelic sites, each subject to two-way mutation, and natural selection against one of the alleles (see Methods for details). We studied three types of demographic models thought to be relevant for human populations: (i) a bottleneck; (ii) exponential growth starting from a constant-sized

population; and (iii) a complex demographic model for African Americans (including rapid recent growth) and European Americans (including two bottlenecks followed by growth) inferred by Tennessen *et al.*[5]. The main features of the Tennessen model are similar to other recent models[9, 10, 17] while using a larger data set for parameter estimation. Our main results focus on selection against semi-dominant (i.e., additive) alleles in which the three genotypes have fitnesses 1, $1 - s/2$ and $1 - s$, respectively; and selection against recessive alleles with genotype fitnesses 1, 1, and $1 - s$. The effects of demography in these two models are qualitatively representative of those over the range of dominance coefficients (Appendix 1, Section 2.4). In addition to simulation results shown here, further results and detailed theoretical analysis for all our key results are provided in Appendix 1.

**The impact of demographic changes on individual load.** We focus first on the impact of demographic changes on individual load – that is, we want to understand whether demographic history has impacted the burden of deleterious variation carried by a typical individual in a population. Individual load is directly related to the number of deleterious alleles carried by an individual, or for recessive mutations to the number of homozygous sites per individual (see the Methods and Appendix 1 for further details).

Figure 1.1 illustrates the impact of a bottleneck and population growth on the numbers of deleterious variants with strong selection (s=1%). As expected, these demographic events have a major impact on the number and frequency spectra of deleterious variants: the bottleneck causes a decrease in the total number of segregating sites in a population due largely to loss of rare variants, while the mean frequency of alleles that survive increases. Meanwhile, exponential

growth causes a rapid increase in the number of segregating sites due to a major influx of rare variants, but a consequent drop in the mean frequency at segregating sites. But despite these dramatic shifts in the overall frequency spectrum, the impact on genetic load – namely, the mean number of deleterious variants per individual and thus the average fitness – is much more subtle.

In the semi-dominant case, the load is essentially unaffected by these demographic events (Figures 1.1C and 1.1D). With growth, the increased number of segregating sites is exactly balanced by a decrease in mean frequency (and conversely for the bottleneck), so that the number of variants per individual stays constant. This kind of balance is predicted by classic mutation-selection balance models[18], and can be shown to hold for general changes in population size, provided that selection is strong and deleterious alleles are at least partially dominant (Appendix 1, Section 2.3).

The behavior of the recessive model is more complicated (Figures 1.1E and 1.1F). In the bottleneck model, the mean number of deleterious variants per individual drops by 60% as a result of the bottleneck. This is due to the loss of rare alleles. However, during the bottleneck, some deleterious alleles drift to higher frequencies[11, 19], contributing disproportionately to the number of homozygotes. This causes a transient increase in the number of deleterious homozygous sites per individual – i.e., the recessive load. Meanwhile, population growth has a less pronounced effect on recessive variation, leaving the mean number of deleterious alleles per individual unchanged, but causing a slight decrease in load.

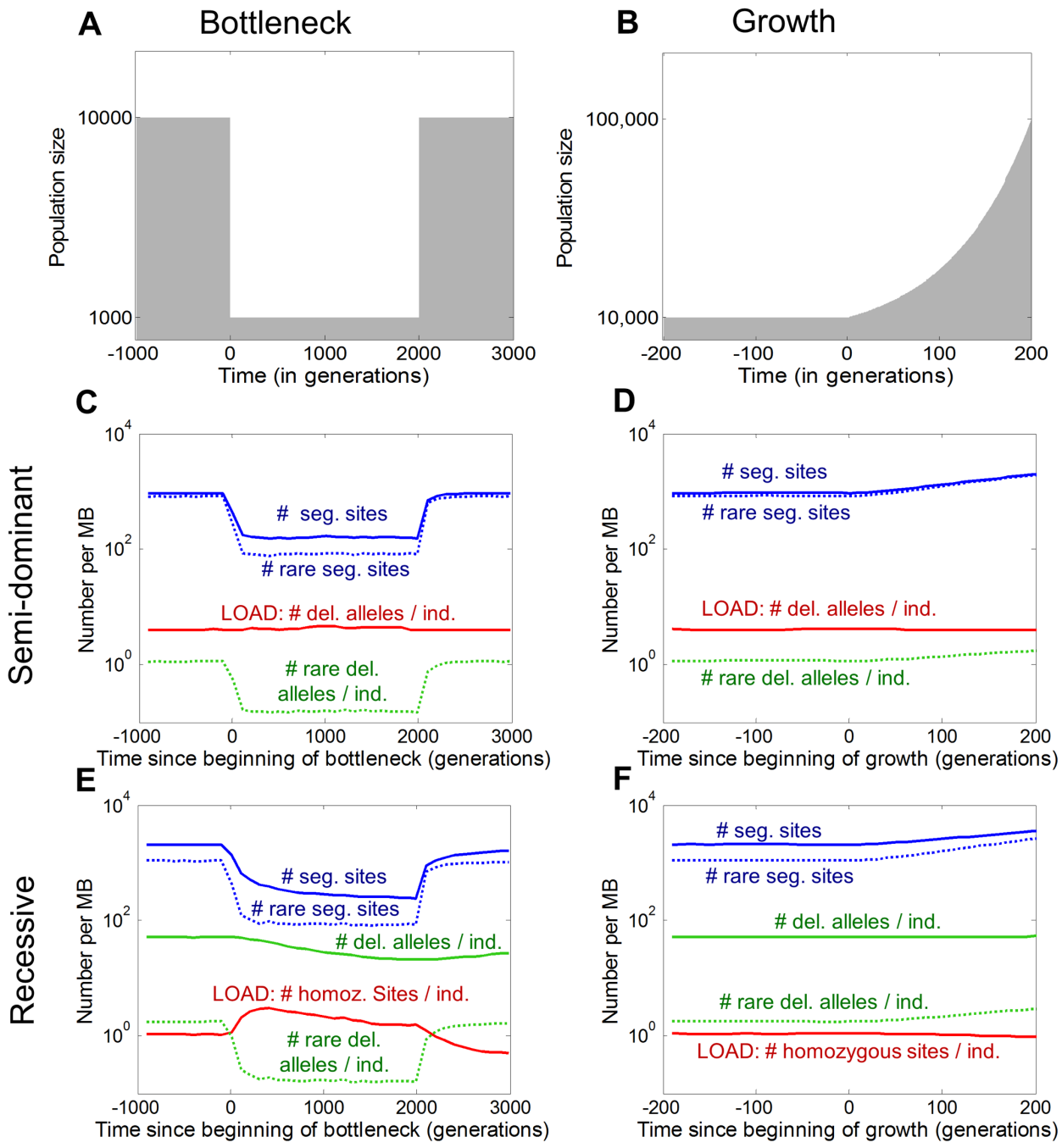



**Figure 1.1. Time course of load and other key aspects of variation through a bottleneck (A) and exponential growth (B).** Each data line shows the expected number of variants, or alleles per MB, assuming semi-dominant mutations (C and D) or recessive mutations (E and F) with $s$ = 1% and mutation rate per site per generation=$10^{-8}$. Versions of these plots with linear scales can be found in Figures A1.13, A1.14, and A1.16.

More generally, the manner in which demography affects load varies with the degree of dominance and the strength of selection (Figure 1.2, Appendix 1, Section 2 & Table A1.1). The behavior of these models can be classified into three selection regimes (strong, weak and effectively neutral). In the strong selection case, i.e., where selection is much stronger than drift (approximately $s \geq 10^{-3}$ for semi-dominant mutations), deleterious variants are extremely unlikely to fix, and virtually all of the genetic load is due to segregating variation. In this range, we infer that human demography has had no impact on semi-dominant load (and more generally for mutations with at least some dominance component), and small effects on recessive load.

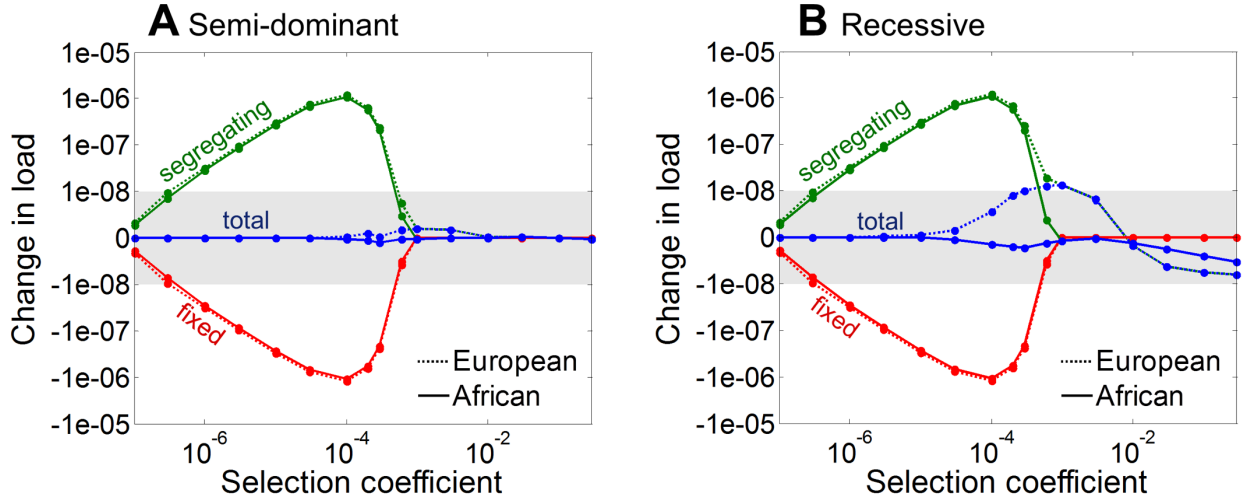



**Figure 1.2. Changes in load due to changes in population size during the histories of European and African Americans for (A) semi-dominant and (B) recessive sites.** The blue lines denote the difference in load per base pair of DNA sequence in the present day population compared to the ancestral (constant) population size, as a function of selection coefficient. The green and red lines show the difference in load due to segregating and fixed variants, respectively. As can be seen, the increase in load due to segregating variation in modern populations approximately cancels out with the decrease in load due to fixed sites. The scale on the y-axis is linear within the grey region and logarithmic outside.

The weak selection case – where drift and selection have comparable effects – is more complex, as fixed alleles may contribute appreciably to load, and steady state load depends on population size[20]. However, the approach to steady state is very slow, being limited both by the time to fixation (on the order of $4N$ generations) and by the mutational input (on the order of $1/2NU$ generations). For both the semi-dominant and recessive cases, population growth is too recent to have substantially decreased the load. Recent growth increases the input of new deleterious mutations, but this effect is counterbalanced by the fact that the new deleterious mutations are proportionally rarer. The bottleneck in Europeans is estimated to have occurred farther in the past and at much lower population sizes[5] (Figure A1.1), allowing it to have more effect. In this case, the increase in drift causes segregating deleterious alleles to increase in frequency, sometimes reaching fixation, and results in a slight increase in load (Figure A1.2). The out-of-Africa bottleneck should thus lead to a slight increase of load in Europeans, most notably for recessive sites.

Finally, in the effectively neutral range – where selection has negligible effects on the population dynamics – segregating variation contributes negligibly and hence the load does not change with demography. Thus, across all three selection regimes, recent human demographic history is likely to have had virtually no impact on genetic load at partially dominant sites, and only weak effects at recessive sites.

**Analysis of exome data.** To test these predictions, we analyzed two recent data sets of exome sequences from individuals of west African and European descent. Previous work comparing load in different populations has produced conflicting conclusions depending on the dataset, choice of

measures and functional annotations. For example, Lohmueller *et al*.[11] reported that there is "proportionally more deleterious variation in European than in African populations". Similarly, Tennessen *et al*.[5] found that European Americans had more non-reference genotypes when they used a conservative classification of deleterious sites, but observed the opposite when using a more liberal classification of sites (both observations were highly significant).

We first analyzed single nucleotide variant (SNV) frequency data from a recent exome sequencing study of 2,217 African Americans (AAs) and 4,298 European Americans (EAs) sequenced at 15,336 protein coding genes by Fu *et al*.[6] (allele frequencies available from the NHLBI GO Exome Variant Server). Additionally, we analyzed exome data from 88 Yoruba (YRI) and 81 European (CEU) individuals collected by the 1000 Genomes Project[21].

To test whether there are differences in load between individuals of west African and European descent, we considered the average number of derived alleles per individual at putatively deleterious segregating sites. For this purpose, a site is considered to be segregating if and only if it is variable within the combined sample of both populations. This definition ensures that the derived counts are comparable across populations. Under a semi-dominant model, the number of derived alleles increases monotonically with the segregating genetic load. Thus, any difference in average load between populations would be apparent as a difference in the mean number of derived alleles per individual. Here, we focused on an equivalent measure that also facilitates comparisons across different types of sites: namely, the mean derived allele frequency within functional classes. Note that the mean derived allele frequency is simply equal to the number of derived alleles per individual divided by twice the number of segregating sites in that

class, and so any difference in the mean number of derived alleles per individual will also be a difference in mean derived frequencies. For sites that are either neutral or semi-dominant, our model predicts that the mean derived allele frequency should be virtually identical in Africans and Europeans (Appendix 1, Section 3 & Figure A1.3). At recessive sites, we expect a slight increase in mean derived frequency in Africans compared to Europeans (Figure A1.3), but overall we expect any differences to be small.

Functional predictions of SNVs were obtained from PolyPhen2, a method that uses sequence conservation and structural information to infer which non-synonymous changes are most likely to have functional consequences[22]; see Table A1.2 for similar analyses with other functional prediction methods. When using the functional predictions we observed a strong bias: SNVs where the genome reference carries the derived allele are much more likely to be classified as benign than SNVs where the reference allele is ancestral—this is true even when we control for the overall population frequency (Figure A1.4). Hence our analysis incorporates a correction to account for this bias; we also obtained very similar results using a separate set of unpublished human-independent PolyPhen scores kindly provided by the Sunyaev lab (Table A1.4).

Figure 1.3 summarizes the results for the data of Fu *et al*. As expected, the mean allele frequency declines with increasing functional severity[5], from 2.8% at noncoding SNVs to 0.6% at probably-damaging SNVs, implying that there is selection against most SNVs with predicted damaging effects. More striking, however, is that within each of the five functional categories, the mean allele frequencies – and hence the numbers of derived alleles per individual – are

essentially identical in the two populations, despite the very large size of the data sets (p>.05 for all five comparisons). Results for the 1000 Genomes Project data are qualitatively similar: we find no significant differences between YRI and CEU in the numbers of derived alleles per individual in any functional category (Table A1.5).

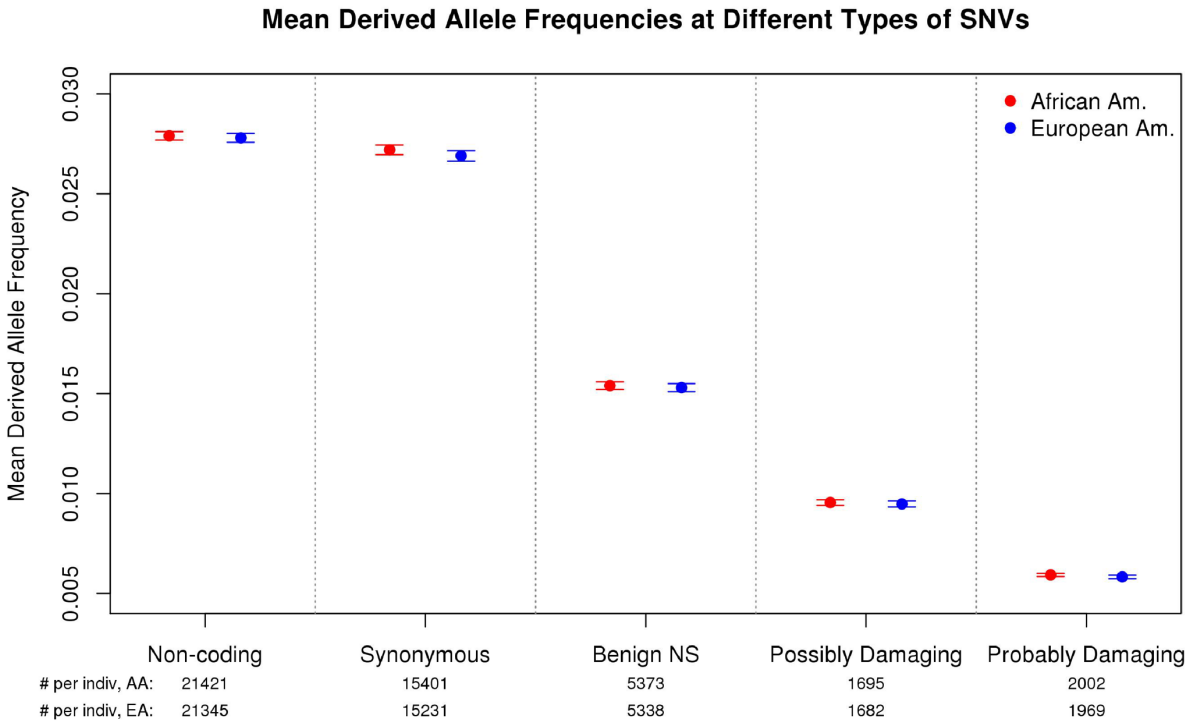



**Figure 1.3. Observed mean allele frequencies in African and European Americans at various classes of SNVs.** The plot shows mean frequencies in each population, plus and minus two standard errors, using exome sequence data from Fu et al.[6]. Here a site is considered an SNV if it is segregating in the combined AA-EA sample of 6515 individuals. The functional classifications of sites are from PolyPhen2[22] with bias-correcting modifications. The AA and EA mean frequencies are essentially identical within all five functional categories ($p$>0.05).

In summary, these observations are consistent with our model predictions that load should be very similar in these populations. Our conclusions likely differ from previous studies partly because earlier studies used measures that are related to load but are also sensitive to other differences between the populations being compared (e.g., the number of neutral segregating

sites and the frequency spectrum) and partly due to the reference bias in functional annotations accounted for here (see Appendix 1, Section 3). We note that David Reich, Shamil Sunyaev and colleagues have recently made similar observations regarding load in different populations (personal communication).

**The impact of demography on the genetic architecture of disease susceptibility.** Although population size changes have had little impact on the average load carried by individuals, growth has greatly increased the number of rare variants in populations. So do rare variants play a greater (and substantial) role in the genetics of disease as a result of recent growth (Figure 1.4)? Given the differences in population history, do higher frequency variants play a greater role in Europeans and Asians than in Africans? The answers to these questions are of practical importance because different study designs may be needed to identify rare variants[13, 15, 16, 23].

To study this, we computed the contributions of different allele frequencies to the heritable phenotypic variation among individuals in the population, namely $x(1-x)f(x)/2$, where $f(x)$ is the probability that a derived allele is at frequency $x$ given the demographic model and selection coefficient. These distributions show the fraction of genetic variance for a disease that is contributed by alleles below frequency x, for the simplest case where the loci underlying a trait all have the same effect size, the same selection coefficient, and are semi-dominant (see Appendix 1, Section 4). In practice, we anticipate that variants underlying a given disease would have a variety of selection coefficients and effect sizes, in which case the overall distribution would be an appropriately weighted mixture of distributions for different selection coefficients. Note that in this model, we consider the proportional contribution of variants at different

frequencies and thus, these results should hold regardless of the number of loci underlying variation in the trait.

Analysis of this model reveals several interesting points. For effectively neutral, or for weakly deleterious sites (Figure 1.4A), only a small fraction of the total variance comes from very rare alleles: although there are many rare alleles, each one contributes very little to population variance and individual load. The same is true for recessive variation across almost the entire range of selection coefficients (Appendix 1, Section 4.2 & Figure A1.5). Likewise, if we assume that the frequency density $f(x)$ follows the frequency spectrum observed at all non-synonymous sites classified as "probably damaging"[22] then, under the same model, it is still only a modest fraction of the genetic variance that is due to rare alleles (Figure 1.4B; cf. ref. [5]). Meanwhile, in all of these cases, the Out-of-Africa bottleneck increases the contribution of intermediate frequency alleles to the genetic variance (Figure 1.4A-C): e.g., at probably damaging sites 62% of the variance in EAs is contributed by alleles with minor allele frequency above 10% compared to only 49% in AAs.

It is only for the case of strong, dominant selection that very rare variants ($< 0.1\%$) become important (Figure 1.4C and 1.4D). For example, for a selection coefficient of 1%, most of the variation is due to rare alleles that arose within the recent exponential growth phase. As a result, the contribution of extremely rare variants is much greater than it would have been in the absence of growth: e.g., in AAs and EAs, 80%, and 65% of the variance is due to alleles below frequency 0.1%, compared to just 25% in the constant population model.

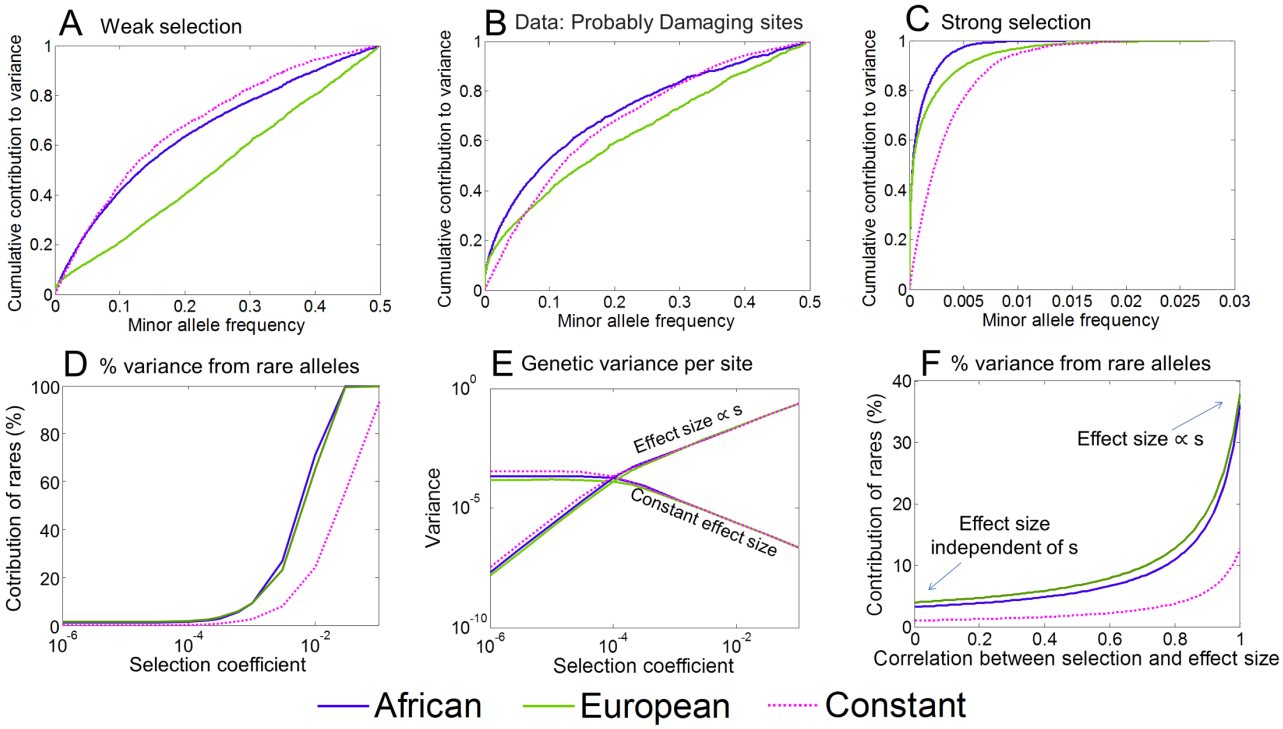



**Figure 1.4. Predicted effect of demography on the genetic architecture of disease risk.** All the plots assume an additive trait and, with the exception of (B), are based on simulations with semi-dominant selection under the Tennessen *et al.*[5] demographic model. Results for the constant population size model are also provided for comparison. The upper plots show the cumulative fractions of genetic variance due to alleles at frequency < *x*, based on: (A) simulated data with weak selection ($s$ =.0002); (B) assuming the observed frequency spectrum at "probably damaging" sites[6, 22], where a constant population size of 14,474 and selection coefficient of 0.02% are used for comparison; and (C) simulated data with strong selection ($s$ = .01). Panel (D) depicts the fraction of variance due to rare alleles (i.e., < 0.1%) as a function of the selection coefficient; (E) shows the per-site contribution to variance as a function of the selection coefficient under two extreme models, with effect sizes that are either independent of $s$ (constant) or proportional to $s$; (F) shows the expected fraction of the variance due to rare variants (i.e., < 0.1%) as a function of the correlation between the selection on, and effect size of variants. Further details on the model are provided in the Methods.

Of course in practice, the genetic variants that contribute to a complex trait likely have a range of selection coefficients ($s$) and a range of effect sizes ($a$) on the phenotype in question (Appendix 1, Section 4.3). When there is a mixture of selective coefficients and effect sizes, what can we say about the relative importance of rare and common variants? The answer crucially depends on the relationship between $a$ and $s$[14, 24]. To illustrate this, we consider two

extreme cases: (1) $a$ is independent of $s$, namely, the trait itself has little effect on fitness but specific variants could have fitness consequences due to pleiotropic effects on other phenotypes; and (2) $a$ is proportional to $s$ — likely most relevant for traits with a direct impact on fitness such as early-onset diseases or diseases affecting fertility. Figure 1.4E shows the expected contribution of each site to genetic variance as a function of $s$, under these two models. When $a$ is independent of $s$, we would expect weakly selected mutations to contribute most of the variance because they have the same average effect on the trait but can drift to higher frequencies. But the reverse occurs in the model where $a$ increases with $s$: highly deleterious, rare mutations will have a greater contribution to variance because their increased effect size outweighs their lower frequencies.

Many traits presumably lie between these two extreme cases. To study how demography affects genetic architecture across this range, we consider a second model. We assume that the heritable variance in a trait is due to a mixture of weakly ($s = 0.0002$) and strongly ($s = 0.01$) selected mutations and we vary the correlation between selection on a variant and its effect on the trait (see Methods for details). Figure 1.4F shows how the contribution of rare alleles to genetic variance changes with the correlation between the selection coefficient and effect size. As can be seen in the case with constant population size, the contribution of rare variants becomes substantial only when the variants' effects on fitness and on the trait are highly correlated (presumably because the trait itself is strongly coupled with fitness). While growth affects the frequencies of strongly selected alleles regardless of the correlation, it will have a substantial effect on the genetic architecture of a trait only for traits in which strongly selected alleles contribute substantially to variance. In this case, we see that the recent growth greatly

amplifies the contribution of rare alleles to the variance. A similar argument implies that the Out-of-Africa bottleneck should substantially increase the contribution of intermediate frequency alleles to the variance, unless the effects of variants on fitness and on the trait are highly correlated, in which case rare alleles will still dominate.

## Conclusion

While recent demographic events have had well-documented effects on the frequency spectrum of SNVs in modern populations, we find that these events have had negligible impact on the average burden of mutations carried by individuals. Moreover, we conclude that although there are large absolute numbers of rare variants, they do not necessarily contribute a large fraction of the genetic variance underlying complex traits. An earlier paper from one of the present authors (Pritchard, 2001[13]) also discussed the possible role of allelic heterogeneity and rare variants in disease using a model that is closer to the independent $s$ model here. While the earlier model is not exactly comparable to our present work, the overall results are broadly consistent, as the bulk of the genetic variance was predicted to be due to variants that would not be considered rare by modern standards. To summarize, it is only for diseases that are primarily due to strongly deleterious mutations that we can expect much of the variance to be due to rare alleles: these will likely tend to be diseases that are tightly coupled to fitness.

## Methods

This section provides a summary of our methods; a complete description may be found in Appendix 1.

**Model.** Our basic model starts by considering selection at a single site. We use the standard bi-allelic diploid model with two-way mutation, viability selection, drift and, in some cases, migration[25]. Specifically, we assume there are two possible alleles at each site: normal (N) and deleterious (D). An N allele mutates to the D allele with probability u per gamete, per generation and the reverse mutation occurs with probability v. Unless noted otherwise, we assume that mutation is symmetric, i.e., u=v. The absolute fitness of the three genotypes NN, ND and DD are 1, $1 - hs$ and $1 - s$, respectively, where $s > 0$ and $h \geq 0$. We focus on semi-dominant ($h = 1/2$) and fully recessive ($h = 0$) selection because these two cases exhibit the full range of qualitative behaviors, with selection acting primarily on heterozygotes when h>1/2 and only on homozygotes when h=0. Allele frequencies in the next generation follow from Wright-Fisher sampling with these viabilities, sometimes with migration, and the population size and migration rates vary according to the demographic scenario considered.

We assume that fitness is multiplicative across sites, and that there is linkage equilibrium among sites. Under these assumptions, the evolutionary dynamics at each site are independent from all other sites. In practice, linked selection is likely to have negligible effects on differences between populations because, to a first approximation this reduces the effective population size at a given site by similar proportions regardless of demographic history and these effects are thought to be modest in humans (e.g., ref. [26]).

**Demographic scenarios.** We consider three demographic scenarios. The most detailed is the Out-of-Africa demographic model for African-Americans (AA) and European-Americans (EA) estimated by Tennessen *et al*.[5] (Figure A1.1A). The model includes the Out-of-Africa split of European ancestors, changes in population size before and after the split (specifically, a severe bottleneck in Europeans following the split and recent rapid growth in both Europeans and Africans) and migration between the populations after the split. Finally, the model includes recent admixture between the populations, which we include in our simulations only when we compare our results to data from AAs.

We also study two simpler demographic scenarios (Figure A1.1B&C). To understand the effects of recent explosive growth of human populations, we use a simple model of exponential growth from a population of constant size and similarly, to investigate the effects of the bottleneck in Europeans at the Out-of-Africa split, we consider a simple model of a bottleneck where population size instantaneously changes to a lower value at which it stays constant until it instantaneously reverts back to its original size.

**Simulations.** For each demographic scenario, we run simulations of a single site for the semi-dominant and recessive cases and vary the selection coefficient such that the strength of selection ranges from effectively neutral to strong. Each run begins with one of the two alleles fixed, where the proportion of runs that start with each allele is given by the expectation at equilibrium. A burn-in period of $\geq 10N$ generations with constant population size $N$ follows in order to ensure an equilibrium distribution of segregating sites. The initial state is defined as ancestral and the other state as derived; the derived and deleterious allele frequencies are

recorded at the end of the simulation. The code is written in C++ and is available upon request. (See Appendix 1, Section 1 & Figures A1.6-A1.8.)

**Load.** Genetic load is defined as the relative reduction in average fitness caused by deleterious alleles, compared to the maximum absolute fitness [25]. In our model, the maximal absolute fitness equals 1, allowing us to directly consider differences in average fitness in populations with different demographic histories. Given our model, the average fitness function can be written as

$$\bar{W} \approx \exp\left(-\sum_{j=1}^{M} l(h_j, s_j)\right)$$

where

$$l(h,s) \equiv 2hsE(pq) + sE(q^2) = s\big(2hE(q) + (1-2h)E(q^2)\big), \qquad (1.1)$$

relates the quantities at a locus with load, $p$ and $q$ are the beneficial and deleterious allele frequencies at a locus ($p + q = 1$) and $h_j$ and $s_j$ are the dominance and selection coefficient at locus $j$. For a model with a single site and $s \ll 1$, $l(h,s)$ coincides with the definition of load. For more than one site, load is a simple function of the sum over $l(h, s)$'s. For brevity, we therefore refer to $l(h,s)$ as load.

**Change in load.** To assess whether there has been a change in load due to demography, we consider the difference between load at the present time and the load before recent demographic events. Specifically, in the exponential and bottleneck models the reference time is before the change in population size and in the Tennessen model the reference time is the split between the

African and European populations. (See Appendix 1, Section 2, Figures A1.2, A1.9-20 & Table A1.1.)

**Data Analysis.** We used exome resequencing data from Fu *et al*. (2012)[6] and from the 1000 Genomes Project[21]. Allele frequency estimates from Fu *et al*. are available from the NHLBI GO Exome Variant Server. These provide estimates of the derived allele frequencies (DAFs) at exonic SNVs in European-Americans (EAs) and African-Americans (AAs). 1000 Genomes Project vcf files (Phase 1, Version 3) were downloaded from the official 1000 Genomes public server. YRI and CEU individuals with (at least) exome sequencing coverage were extracted from the original .vcf files (88 YRI individuals and 81 CEU individuals). 7 YRI individuals, chosen at random, were removed to match sample sizes between YRI and CEU. Variants that were fixed for either allele in both populations were removed. Any variant that was not an SNV or did not contain ancestral allele information was also dropped.

The ANNOVAR suite of scripts [27] was used to obtain functional predictions for each SNP from each of four prediction methods: PolyPhen2[22], SIFT[28], LRT[29] and MutationTaster[30]. We observed a strong reference bias in the functional classifications for all four prediction methods: sites at which the reference genome carries the derived allele are much more likely to be classified as benign than are sites where the reference is ancestral; this is a very strong effect even when we control for the true population frequency in a very large sample (Figure A1.4), and hence does not simply reflect the tendency for common alleles to be less functional. We therefore treated the functional designations at sites where the genome reference is derived as unreliable. To deal with this problem we used a simple procedure to estimate the probability that

each reference-derived site would have been classified as damaging had the reference allele been ancestral (conditional on the overall population frequency). Specifically, we binned SNVs by overall population frequency in the full sample and, for each bin, we determined the fraction of reference-ancestral sites in each functional category. For SNVs in that bin that are reference-derived, we treated those fractions as estimates of the probability that these SNVs *would* have been in each functional category had they instead been reference-ancestral. Next, to estimate the mean derived allele frequency (DAF) for each functional category, we summed across all sites in that category that were reference ancestral, and added a contribution from all sites that were reference-derived, weighted according to the estimated probability that the site would have been in the relevant functional category if it had been reference-ancestral. We also provide supplementary results in which we used a new unpublished version of PolyPhen's PSIC scores that are calculated in a human-independent (i.e., unbiased) manner and obtain qualitatively similar results. We thank Ivan Adzhubey and Shamil Sunyaev for pre-publication access to these.

We calculated mean derived frequencies within functional categories, and the corresponding standard errors (calculated as SD(DAF)/Sqrt(#sites)). Individual-level counts for the 1000 Genomes data simply counted the numbers of derived alleles per individual within a functional class (note that there are no missing genotypes in this data set as these have been imputed by the 1000 Genomes Project). For each population and functional category we estimated the standard deviation of the mean number of derived alleles per individual by bootstrapping across sites. This is more appropriate than computing the standard error directly from the distribution of derived allele counts across individuals, as the latter method ignores variation in the

evolutionary process. Note that because we are working with mean allele counts or frequencies, these analyses are unaffected by linkage disequilibrium or Hardy Weinberg disequilibrium (which may affect variances but not means).

Note that our analysis effectively uses the derived allele count as a proxy for the deleterious allele count. Hence, there will be a low rate of misclassification at weakly selected sites for which the deleterious allele is ancestral. However this does not change the qualitative predictions about patterns of differences between populations and we expect the number of derived alleles to have a monotonic relationship with the number of deleterious alleles. Specifically, for sites that are either neutral or semi-dominant, we predict that this measure should yield virtually identical counts in AAs and EAs (Appendix 1, Section 2 & Figure A1.20). At recessive sites, our model predicts slight differences (Appendix 1, Section 2), but overall we expect these differences to be negligibly small. Note that when SNVs are defined within populations as in some previous papers, these simple predictions do not hold.

**Models for variance.** We consider how the relationship between the effects of mutations on fitness and a trait affect genetic architecture. For that purpose, we calculate the expected contribution of mutations to the heritable variation in a trait. We assume an additive trait and that the fitness effects of mutations are semi-dominant. At a site with selection coefficient $s$, the expected contribution to the variance from deleterious alleles below frequency $\omega$ is therefore

$$V_\omega(s) = \frac{1}{2} CE(a^2|s) \int_0^\omega f(x|s)x(1-x)dx\,, \tag{1.2}$$

where $E(a^2|s)$ is the expectation of the squared effect size, $f(x|s)$ is the probability of the deleterious allele being at frequency $x$ (without conditioning of the site being segregating, i.e.,

including $x$ = 0 and 1) and the $C$ is a proportion coefficient (cf. Appendix 1, Section 4.1). A site's expected contribution to variance is $V_1(s)$and the proportional contribution from variants below frequency $\omega$ is $\Theta_\omega(s) \equiv \frac{V_\omega(s)}{V_1(s)}$; Note that while $V_1(s)$ depends on the relationship between selection coefficients and effect sizes, $\Theta_\omega(s)$ does not. When all sites are considered jointly, denoting the input of mutations with selection coefficient $s$ by $\mu(s)$, the expected proportion of variance from deleterious alleles below frequency ω is

$$\Theta_\omega = \frac{\int_s \mu(s) V_1(s)\, \Theta_\omega(s)\, ds}{\int_s \mu(s) V_1(s)\, ds}. \qquad (1.3)$$

As an illustration, we consider a simple model in which we vary the correlation between selection on variants and their effects on a trait. We assume that half of the newly arising mutations have a weak selection coefficient $s_w$ = 0.0002 and half have a strong selection coefficient of $s_s$ = 0.01. For strongly selected mutations, the effect size on the trait, $a$, is chosen to be $cs_s$ with probability $\frac{1}{2}(1 + p)$ and $cs_w$ with probability $\frac{1}{2}(1 - p)$, where $c$ is a positive constant and $0 \leq p \leq 1$; correspondingly, for weakly selected mutations the effect size is chosen to be $cs_w$ with probability $\frac{1}{2}(1 + p)$ and $cs_s$ with probability $\frac{1}{2}(1 - p)$. In this model, the marginal distributions of selection coefficients and effect sizes do not depend on $p$, while the correlation between them is equal to $p$. To obtain Figure 1.4F we therefore varied $p$ between 0 and 1. In Figure 1.4E, we consider the two extremes ($p$ = 0 and 1).

**URLs.** The NHLBI GO Exome Variant Server, http://evs.gs.washington.edu/EVS; The 1000 Genomes public server, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/.

**Acknowledgements**

This work was supported by grants from the National Institutes of Health (MH084703, GM083228), the Israel Science Foundation (grant # 1492/10), and the Howard Hughes Medical Institute. MCT was supported in part by NIH grant T32 GM007197. Thanks to David Reich and Shamil Sunyaev for helpful discussions and generous input regarding the interpretation of PolyPhen 2; to Ivan Adzhubey for human-independent PolyPhen scores; to Josh Akey for assistance in accessing data; and to Josh Akey, Adam Siepel, Graham Coop, Ilan Eshel, Dick Hudson and two anonymous reviewers for comments on the manuscript; and to Molly Przeworski for many discussions and comments on the manuscript.

# Chapter 2
# A population genetic interpretation of GWAS findings for human quantitative traits

(Published under Simons et al. PLOS Biology 2018)

## Abstract

**Human genome-wide association studies (GWAS) are revealing the genetic architecture of anthropomorphic and biomedical traits, i.e., the frequencies and effect sizes of variants that contribute to heritable variation in a trait. To interpret these findings, we need to understand how genetic architecture is shaped by basic population genetics processes—notably, by mutation, natural selection and genetic drift. Because many quantitative traits are subject to stabilizing selection and genetic variation that affects one trait often affects many others, we model the genetic architecture of a focal trait that arises under stabilizing selection in a multi-dimensional trait space. We solve the model for the phenotypic distribution and allelic dynamics at steady state and derive robust, closed form solutions for summary statistics of the genetic architecture. Our results provide a simple interpretation for missing heritability and why it varies among traits. They predict that the distribution of variances contributed by loci identified in GWAS is well approximated by a simple functional form that depends on a single parameter: the expected contribution to genetic variance of a strongly selected site affecting the trait. We test this prediction against the results of GWAS for height and body mass index (BMI) and find that it fits the data well, allowing us to make inferences about the degree of pleiotropy and mutational target size for these traits. Our findings help to explain why the GWAS for height explains more of the heritable variance than similarly-sized GWAS for BMI, and to predict the increase in explained heritability with study sample size. Considering the demographic history of European populations, in which these GWAS were performed, we further find that most of the associations they identified likely involve mutations that arose during the out of Africa bottleneck at sites with selection coefficients around $s = 10^{-3}$.**

# Introduction

Much of the phenotypic variation in human populations, including variation in morphological, life history and biomedical traits, is "quantitative", in the sense that heritable variation in the trait is largely due to small contributions from many genetic variants segregating in the population [1,2]. Quantitative traits have been studied since the birth of biometrics over a century ago [1-3], but only in the past decades have technological advances made it possible to systematically dissect their genetic basis [4-6]. Notably, since 2007, genome-wide association studies (GWAS) in humans have led to the identification of many thousands of variants reproducibly associated with hundreds of quantitative traits, including susceptibility to a wide variety of diseases [4]. While still ongoing, these studies already provide important insights into the genetic architecture of quantitative traits, i.e., the number of variants that contribute to heritable variation, as well as their frequencies and effect sizes.

Perhaps the most striking observation to emerge from these studies is that, despite the large sample size of many GWAS, all variants significantly associated with any given trait typically account for less (often much less) than 25% of the narrow sense heritability [4,7,8,]but see [9]. (Henceforth, we use "heritability" to refer to narrow sense heritability.) While many factors have been hypothesized to contribute to the "missing heritability" [7,8,10-14], the most straightforward explanation and the emerging consensus is that much of the heritable variation derives from variants with frequencies that are too low or effect sizes that are too small for current studies to detect. Comparisons among traits also suggest that there are substantial differences in architectures. For example, recent meta-analyses GWAS uncovered seven times as many variants for height (697) than for body mass index (97), and together the variants for height

account for more than four times the heritable variance for body mass index (~20% vs. ~3-5%, respectively), despite comparable sample sizes [15,16].

These first glimpses underscore the need for theory that relates the findings emerging from GWAS with the evolutionary processes that shape genetic architectures. Such theory would help to interpret the "missing heritability" [17-20] and to explain why architecture differs among traits. It may also allow us to use GWAS findings to make inferences about underlying evolutionary parameters, helping to answer enduring questions about the processes that maintain phenotypic variation in quantitative traits [5,21].

Development of such theory can be guided by empirical observations and first principles considerations. New mutations affecting a trait arise at a rate that depends on its "mutational target size" (i.e., the number of sites at which a mutation would affect the trait). Once they arise, the trajectories of variants through the population are determined by the interplay between genetic drift, demographic processes, and natural selection acting on them. These processes determine the number and frequencies of segregating variants underlying variation in the trait. The genetic architecture further depends on the relationship between the selection on variants and their effects on the trait. Notably, selection on variants depends not only on their effect on the focal trait but also on their pleiotropic effects on other traits. We therefore expect both direct and pleiotropic selection to shape the joint distribution of allele frequencies and effect sizes.

Multiple lines of evidence suggest that many quantitative traits are subject to stabilizing selection, i.e., selection favoring an intermediate trait value [5,22-27]. For instance, a decline in fitness components (e.g., viability and fecundity) is observed with displacement from mean values for a variety of traits in human populations [28-30], in other species in the wild [31,32] and in experimental manipulations [31,33]. While less is known about complex diseases, they may often

reflect large deviations of an underlying continuous trait from an optimal value [1], with these continuous traits subject to directional (purifying) selection in some cases and to stabilizing selection in others. What remains unclear is the extent to which stabilizing selection is acting directly on variation in a given trait or is "apparent", i.e., results from pleiotropic effects of this variation on other traits.

Other lines of evidence suggest that pleiotropy is pervasive. For one, theoretical considerations about the variance in fitness in natural populations and its accompanying genetic load suggest that only a moderate number of independent traits can be effectively selected on at once [34]. Thus, the aforementioned relationships between the value of a focal trait and fitness are likely heavily affected by the pleiotropic effects of genetic variation on other traits [25,34-36]. Second, many of the variants detected in human GWAS have been found to be associated with more than one trait [37-41]. For example, a recent analysis of GWAS revealed that variants that delay the age of menarche in women tend to delay the age of voice drop in men, decrease body mass index, increase adult height, and decrease risk of male pattern baldness [37]. More generally, the extent of pleiotropy revealed by GWAS appears to be increasing rapidly with improvements in power and methodology [37,42-45]. These considerations and others [45,46] point to the general importance of pleiotropic selection on quantitative genetic variation.

The discoveries emerging from human GWAS further suggest that genetic variance is dominated by additive contributions from numerous variants with small effect sizes. Dominance and epistasis may be common among newly arising mutations of large effect e.g., [47,48-51], but both theory and data suggest that they play a minor role in shaping quantitative genetic variation within populations e.g., [9,52,53-56]. Indeed, for many traits, most or all of the heritability explained in GWAS arises from the additive contribution of variants with squared effect sizes that are

substantially smaller than the total genetic variance e.g., [15,16,57,58]. Moreover, statistical quantifications of the total genetic variance tagged by genotyping (i.e., not only due to the genome-wide significant associations) suggest that such contributions may account for most of the heritable variance in many traits e.g., [9,59-61]. Finally, considerable efforts to detect epistatic interactions in human GWAS have, by and large, come up empty-handed [9,56,62], with few counter-examples mostly involving variants in the MHC region [53,56,63,64,] but see [65]. Thus, while the discovery of epistatic interactions may be somewhat limited by statistical power [56], theory and current evidence suggest that that non-additive interactions play a minor role in shaping human quantitative genetic variation. Motivated by these considerations, we model how direct and pleiotropic stabilizing selection shape the genetic architecture of continuous, quantitative traits by considering additive variants with small effects and assuming that together they account for most of the heritable variance.

To date, there has been relatively little theoretical work relating population genetics processes with the results emerging from GWAS. Moreover, the few existing models have reached divergent predictions about genetic architecture, largely because they make different assumptions about the effects of pleiotropy. Focusing on disease susceptibility, Pritchard [19] considered the "purely pleiotropic" extreme, in which selection on variants is independent of their effect on the trait being considered. In this case, we expect the largest contribution to genetic variance in a trait to come from mutations that have large effect sizes but are also weakly selected or neutral, allowing them to ascend to relatively high frequencies. Other studies considered the opposite extreme, in which selection on variants stems entirely from their effect on the trait under consideration [26,66-70], and have shown that the greatest contribution to genetic variance would arise from strongly selected mutations [67,68] (we return to this case below).

In practice, we expect most traits to fall somewhere in between these two extremes. While there are compelling reasons to believe that quantitative genetic variation is highly pleiotropic, the effects of variants on different traits are likely to be correlated. Thus, even if a given trait is not subject to selection, variants that have a large effect on it will also tend to have larger effects on traits that are under selection e.g., by causing large perturbation to pathways that affect multiple traits; [36,45]. Motivated by such considerations, Eyre-Walker (2010), Keightley and Hill (1990), and Caballero et al. (2015) considered models in which the correlation between the strength of selection on an allele and its effect size can vary between the purely pleiotropic and direct selection extremes. These models diverge in their predictions about architecture, however. Assuming, as seems plausible, an intermediate correlation between the strength of selection and effect size, Eyre-Walker finds that genetic variance should be dominated by strongly selected mutations [20], whereas Keightley & Hill and Caballero et al. conclude that the greatest contribution should arise from weakly selected ones [18,71]. Their conclusions differ because of how they chose to model the relationship between selection and effect size, a choice based largely on mathematical convenience. We approach this problem by explicitly modeling stabilizing selection on multiple traits, thereby learning, rather than assuming, the relationship between selection and effect sizes.

## The Model

We model stabilizing selection in a multi-dimensional phenotype space, akin to Fisher's geometric model [72]. An individual's phenotype is a vector in an *n*-dimensional Euclidian space, in which each dimension corresponds to a continuous quantitative trait. We focus on the architecture of one of these traits (say, the 1$^{st}$ dimension), where the total number of traits parameterizes pleiotropy. Fitness is assumed to decline with distance from the optimal phenotype

positioned at the origin, thereby introducing stabilizing selection. Specifically, we assume that absolute fitness takes the form

$$\mathrm{W}(\vec{r}) = \exp\left(-\frac{r^2}{2w^2}\right), \qquad (2.1)$$

where $\vec{r}$ is the ($n$-dimensional) phenotype, $r = \|\vec{r}\|$ is the distance from the origin and $w$ parameterizes the strength of stabilizing selection. However, we later show that the specific form of the fitness function doesn't matter. Moreover, the additive environmental contribution to the phenotype can be absorbed into $w$ [73; Appendix 2, Section 1.1]; we therefore consider only the genetic contribution.

The genetic contribution to the phenotype follows from the multi-dimensional additive model [74]. Specifically, we assume that the number of genomic sites affecting the phenotype (the target size) is very large, $L \gg 1$, and that allelic effects on the phenotype at these sites are vectors in the $n$-dimensional trait space. An individual's phenotype then follows from adding up the effects of her or his alleles, i.e.,

$$\vec{r} = \sum_{l=1}^{L}(\vec{a}_l + \vec{a}'_l), \qquad (2.2)$$

where $\vec{a}_l$ and $\vec{a}'_l$ are the phenotypic effects of the parents' alleles at site $l$.

The population dynamics follows from the standard model of a diploid, panmictic population of constant size $N$, with non-overlapping generations. In each generation, parents are randomly chosen to reproduce with probabilities proportional to their fitness (i.e., Wright-Fisher sampling with viability selection), followed by mutation, free recombination (i.e., no linkage) and Mendelian segregation. We further assume that the mutation rate per site, u, and the population size are sufficiently small that no more than two alleles segregate at any time at each site (i.e., that $\theta = 4Nu \ll 1$) and therefore an infinite sites approximation applies. The number of

mutations per gamete, per generation therefore follows a Poisson distribution with mean $U = Lu$; based on biological considerations (see Appendix 2, Sections 4.1 and 4.2), we also assume that $1 \gg U \gg 1/2N$. The size of mutations in the $n$-dimensional trait space, $a$ $(= \|\vec{a}\|)$, is drawn from some distribution, assuming only that $a^2 \ll w^2$. We later show that this requirement is equivalent to the standard assumption about selection coefficients satisfying $s \ll 1$ (also see Appendix 2, Section 4.3). The directions of mutations are assumed to be isotropic, i.e., uniformly distributed on the hypersphere in $n$-dimensions defined by their size, although we later show that our results are robust to relaxing this assumption as well.

## Results

**The phenotypic distribution.** In the first three sections, we develop the tools that we later use to study genetic architecture. We start by considering the equilibrium distribution of phenotypes in the population and generalizing previous results for the case with a single trait [26,66,67,70]. Under biologically sensible conditions, this distribution is well approximated by a tight multivariate normal centered at the optimum. Namely, the distribution of $n$-dimensional phenotypes, $\vec{r}$, in the population, is well approximated by the probability density function:

$$\mathrm{f}(\vec{r}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right), \qquad (2.3)$$

where $\sigma^2$ is the genetic variance of the phenotypic distances from the optimum (see Eq. A2.25 for closed form); and under plausible assumptions about the rate and size of mutations (i.e., when $1 \gg U \gg 1/2\mathrm{N}$ and $a^2 \ll w^2$), it satisfies $\sigma^2 \ll w^2$, implying small variance in fitness in the population (Appendix 2, Section 4.2). Intuitively, the phenotypic distribution is normal because it derives from additive and (approximately) independently and identically distributed contributions from many segregating sites. Moreover, the population mean remains extremely

close to the optimum because stabilizing selection becomes increasingly stronger with the displacement from it, and because any displacement is rapidly offset by minor changes to allele frequencies at many segregating sites.

With phenotypes close to the optimum, only the curvature of the fitness function at the optimum (i.e., the multi-dimensional second derivative) affects the selection acting on individuals. In addition, it is always possible to choose an orthonormal coordinate system centered at the optimum, in which the trait under consideration varies along the first coordinate and a unit change in other traits (along other coordinates) near the optimum have the same effect on fitness. These considerations suggest that the equilibrium behavior is insensitive to our choice of fitness function around the optimum. Moreover, in Appendix 2 (Section 5), we show that the rapid offset of perturbations of the population mean from the optimum (by minor changes to allele frequencies at numerous sites) lends robustness to the equilibrium dynamics with respect to the presence of major loci, moderate changes in the optimal phenotype over time and moderate asymmetries in the mutational distribution.

**Allelic dynamic.** Next, we consider the dynamic at a segregating site, and generalize previous results for the case with a single trait [68-70]. This dynamic can be described in terms of the first two moments of change in allele frequency in a single generation (see, e.g., [75]). To calculate these moments for an allele with phenotypic effect $\vec{a}$ and frequency $q$ (=1-$p$), we note that the phenotypic distribution can be well approximated as a sum of the expected contribution of the allele to the phenotype, $2q\vec{a}$, and the distribution of contributions to the phenotype from all other sites, $\vec{R}$. From Eq. 2.3, it then follows that the distribution of background contributions is well approximated by probability density:

$$\mathrm{f}\left(\vec{R}\middle|\vec{a},q\right)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\left(\vec{R}+2q\vec{a}\right)^2}{2\sigma^2}\right). \quad (2.4)$$

By averaging the fitness of the three genotypes at the focal site over the distribution of genetic backgrounds, we find that the first moment is well approximated by

$$\mathrm{E}(\Delta q)\approx\frac{a^2}{w^2}pq\left(q-\frac{1}{2}\right), \quad (2.5)$$

assuming that $a^2$ and $\sigma^2 \ll w^2$ (Appendix 2, Section 4). By the same token, we find that

$$\mathrm{V}(\Delta q)\approx\frac{pq}{2N}, \quad (2.6)$$

which is the standard second moment with genetic drift.

The functional form of the first moment is equivalent to that of the standard viability selection model with under-dominance. This result is a hallmark of stabilizing selection on (additive) quantitative traits: with the population mean at the optimum, the dynamics at different sites are decoupled and selection at a given site acts to reduce its contribution to the phenotypic variance ($2a^2pq$), thereby pushing rare alleles to loss. Comparison with the standard viability selection model shows that the selection coefficient in our model is $s=a^2/w^2$, or $S=2Ns=2Na^2/w^2$ in scaled units. In other words, the selection acting on an allele is proportional to its size-squared in the *n*-dimensional trait space (where *w* translates effect size into units of fitness).

**The relationship between selection and effect size.** The statistical relationship between the strength of selection acting on mutations and their effect on a given trait follows from the aforementioned geometric interpretation of selection. Specifically, all mutations with a given selection coefficient, s, lie on a hypersphere in *n*-dimensions with radius $a=2w\sqrt{s}$, and any given mutation satisfies

$$s=\frac{1}{w^2}a^2=\frac{1}{w^2}\sum_{i=1}^{n}a_i^2, \quad (2.7)$$

where $a_i$ is the allele's effect on the i-th trait (Figure 2.1A). Our assumption that mutation is isotropic then implies that the probability density of mutations on the hypersphere is uniform.

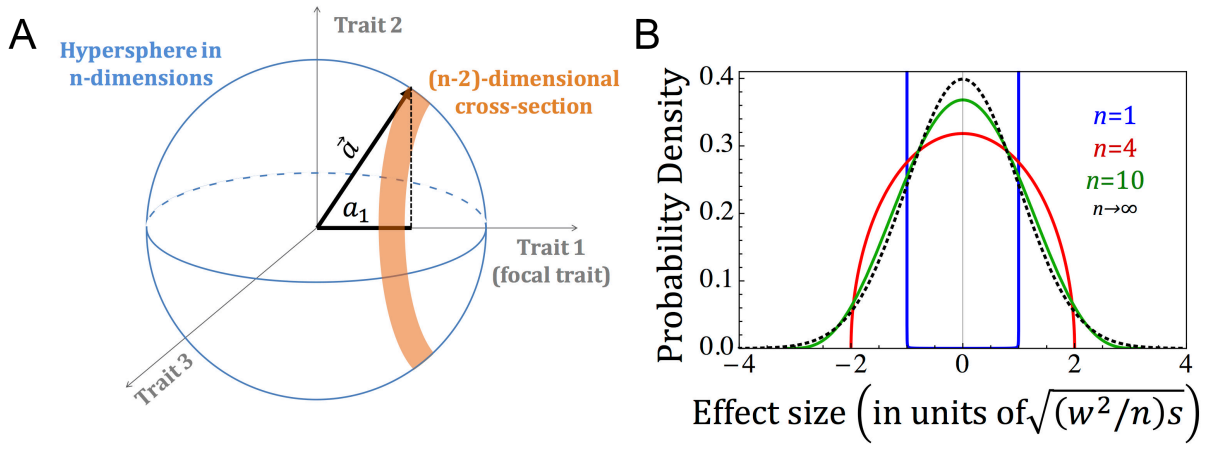



**Figure 2.1. The distribution of effect sizes corresponding to a given selection coefficient.** (A) Mutations with selection coefficient, $\boldsymbol{s}$, lie on a hypersphere in $n$ dimensions with radius $\boldsymbol{a = w\sqrt{s}}$. The probability that such mutations have effect size $\boldsymbol{a_1}$ on the focal trait is proportional to the volume of the $\boldsymbol{(n-2)}$–dimensional cross section of the hypersphere, with projection $\boldsymbol{a_1}$ on the coordinate corresponding to the trait. (B) The distribution of effect sizes on the focal trait, conditional on the selection coefficient being $s$, measured in units of the distribution's standard deviation (see Eq. 2.11).

The distribution of effect sizes on a focal trait, $a_1$, corresponding to a given selection coefficient, $s$, follows. Given that mutation is symmetric in any given trait, $\mathrm{E}(a_1|s) = 0$, and given that it is symmetric among traits,

$$\mathrm{E}(a_1^2|s) = a^2/n = (w^2/n)s. \qquad (2.8)$$

More generally, the probability density corresponding to an effect size $a_1$ is proportional to the volume of the $(n-2)$ – dimensional cross section of the hypersphere with projection $a_1$ (Figure 2.1A). For a single trait, this implies that $a_1 = \pm a$ with probability ½, and for $n > 1$, it implies the probability density

$$\varphi_n(a_1|a) = \frac{\Gamma(n/2)/\Gamma((n-1)/2)}{\sqrt{n/2}} \frac{1}{\sqrt{2\pi(a^2/n)}} \left(1 - \frac{1}{n}\frac{a_1^2}{(a^2/n)}\right)^{\frac{n-3}{2}} \quad (2.9)$$

(Appendix 2, Section 1.2). Intriguingly, when the number of traits $n$ increases, this density approaches a normal distribution, i.e.,

$$\frac{a_1}{\sqrt{a^2/n}} \sim \mathrm{N}(0,1), \quad (2.10)$$

implying that the distribution of effect sizes given the selection coefficient becomes

$$a_1 \sim \mathrm{N}(0, (w^2/n)s). \quad (2.11)$$

This limit is already well approximated for a moderate number of traits (e.g., $n$=10; Figure 2.1B).

The limit behavior also holds when we relax the assumption of isotropic mutation. This generalization is important because, having chosen a parameterization of traits in which the fitness function near the optimum is isotropic, we can no longer assume that the distribution of mutations is also isotropic [76]. Specifically, mutations might tend to have larger effects on some traits than on others, and their effects on different traits might be correlated. In Appendix 2 (Section 5.4), we show that the limit distribution (Eq. 2.11) also holds for anisotropic mutation (excluding pathological cases). To this end, we introduce the concept of an effective number of traits, $n_e$, which can take any real value ≥1, and is defined as the number of equivalent traits required to generate the same relationship between the strength of selection on mutations and their expected effects on the trait under consideration (i.e., replacing $n$ in Eq. 2.11). The robustness of our model, along with mounting evidence that genetic variation is highly pleiotropic (see Introduction), suggest that the limit form may apply quite generally. In that regard, we note that even in this limit, the strength of selection on mutations and their effects on the focal trait are correlated, implying that the kind of "purely pleiotropic" extreme postulated in previous work cannot arise [18-20].

**Genetic architecture.** We can now derive closed forms for summary statistics of the genetic architecture (see Appendix 2, Section 2.3). For mutations with a given selection coefficient, the frequency distribution follows from the diffusion approximation based on the first two moments of change in allele frequency (Eqs. 2.5 and 2.6; [75]), and the distribution of effect sizes follows from the geometric considerations of the previous section. Conditional on the selection coefficient, these distributions are independent and therefore the joint distribution of frequency and effect size equals their product. Summaries of architecture can be expressed as expectations over the joint distribution of frequencies and effect sizes for a given selection coefficient, and then weighted according to the distribution of selection coefficients. While we know little about the distribution of selection coefficients of mutations affecting quantitative traits, we can draw general conclusions from examining how summaries of architecture depend on the strength of selection.

**Expected variance per site.** We focus on the distribution of additive genetic variance among sites, a central feature of architecture that is key to connecting our model with GWAS results. We start by considering how selection affects the expected contribution of a site to additive genetic variance in a focal trait. We include monomorphic sites in the expectation, such that the expected total variance is given by the product of the expectation per-site and the population mutation rate, $2NU$. Under the infinite sites assumption, sites are monomorphic or bi-allelic and their expected contribution to variance is

$$\mathrm{E}(2a_1^2 pq|S) = \mathrm{E}(a_1^2|S)\mathrm{E}(2\,pq|S) = \frac{w^2}{2Nn} S\ \mathrm{E}(2pq|S) \qquad (2.12)$$

(expressed in terms of the scaled selection coefficient *S*). Thus, the degree of pleiotropy only affects the expectation through a multiplicative constant.

This multiplicative factor would have a discernable effect in generalizations of our model in which the degree of pleiotropy varies among sites. For example, if the degree of pleiotropy of one set of sites was $k$ and of another set was $l > k$, and both sets were subject to the same strength of selection, then the expected contribution to genetic variance of sites in the first set would be $l/k$ times greater than in the second (from Eq. 2.12). While such generalizations may prove interesting in the future, here we focus on the model in which the degree of pleiotropy is constant. In this case, the multiplicative factor introduced by pleiotropy is not identifiable from data, because even if we could measure genetic variance in units of fitness (e.g., rather than in units of the total phenotypic variance), we still would not be able to distinguish between the effects of *w* and *n* on the genetic variance per site. We therefore focus on the effect of selection on the relative contribution to variance, which is insensitive to the degree of pleiotropy in our model.

The effect of selection on the relative contribution to genetic variance was described by Keightley and Hill (in the one dimensional case [68]) and is depicted in Figure 2.2B. When selection is strong (roughly corresponding to $S > 30$), its effect on allele frequency (which scales with 1/*S*) is canceled out by its relationship with the effect size (Eq. 2.8), yielding a constant contribution to genetic variance per site, $v_S = 2w^2/nN$, regardless of the selection coefficient (Appendix 2, Section 3.; Figures 2.2A and A2.1B). Henceforth, we measure genetic variance in units of $v_S$. When selection is effectively neutral (roughly corresponding to $S < 1$) and thus too weak to affect allele frequency, the expected contribution of a site to genetic variance scales with the effect size and equals $\frac{1}{2}S\,(\cdot\, v_s)$, and therefore is lower than under strong

selection (Appendix 2, Section 3.1; Figures 2.2A and A2.1A). In between these selection regimes, selection effects on allele frequency are more complex and are influenced by under-dominance (Appendix 2, Section 3.1). As the selection coefficient increases, the expected contribution to variance reaches $v_S$ at $S \approx 3$, and continues to increase until it reaches a maximal contribution that is approximately 30% greater at $S \approx 10$ (Figure 2.2A), after which it slowly declines to the asymptotic value of $v_S$ (Figures 2.2A and A2.1B). Henceforth, we refer to this selection regime as intermediate (not to be confused with the nearly neutral range, which is much narrower and does not include selection coefficients with $S > 10$). These results suggest that effectively neutral sites should contribute much less to genetic variance than intermediate and strongly selected ones [67,68].

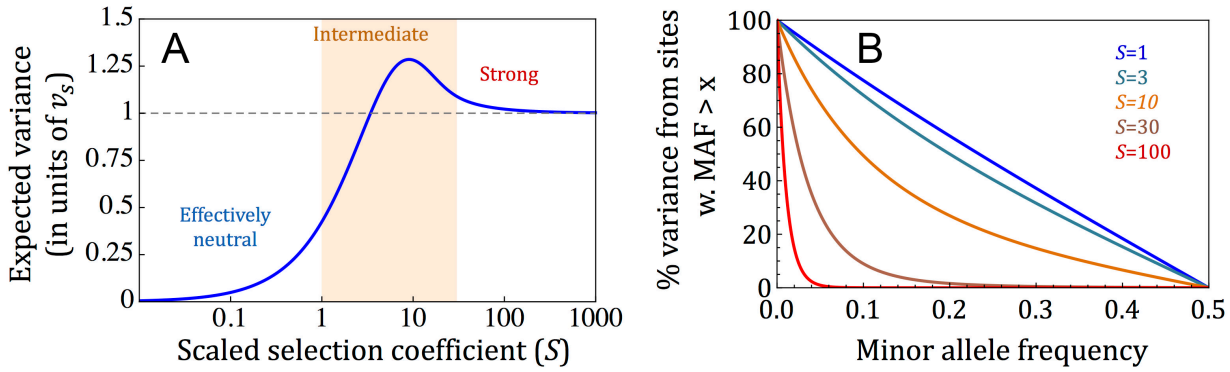



**Figure 2.2. The distribution of additive genetic variance among sites.** In (A), we plot the expected contribution as a function of the scaled selection coefficient. We measure genetic variance in units of $v_S$ – the expected contribution at sites under strong selection. In (B), we show the proportion of additive genetic variance that arises from sites with MAF greater than the value on the x-axis, for different intermediate and strong selection coefficients.

While intermediate and strongly selected sites contribute similarly to variance, their minor allele frequencies (MAF) can differ markedly (Figure 2.2B). As an illustration, segregating sites with MAF $> 0.1$ account for ~72% and ~49% of the additive genetic variance for intermediate selection coefficients of $S$=3 and 10, respectively, when almost no segregating sites would be

found at such high MAF for a strong selection coefficient of $S$=100 (Figure 2.2B). Thus, within the wide range of selection coefficients characterized as intermediate and strong, genetic variance arises from sites segregating at a wide range of MAF ranging from common to exceedingly rare.

**Distribution of variances among sites.** Next, we consider how genetic variance is distributed among sites with a given selection coefficient. We focus on the distribution among segregating sites (including monomorphic effects would just add a point mass at 0). This distribution is especially relevant to interpreting the results of GWAS, because, to a first approximation, a study will detect only sites with contributions to variance exceeding a certain threshold, $v$ ($= 2a_1^2pq$), which decreases as the study size increases (see Discussion). We therefore depict the distribution in terms of the proportion of genetic variance, $\mathrm{G}(v)$, arising from sites whose contribution to genetic variance exceeds a threshold $v$.

We begin with the case without pleiotropy ($n$=1), in which selection on an allele determines its effect size (Figure 2.3A). When selection is strong ($S$>30), the proportion of genetic variance exceeding a threshold $v$ is also insensitive to the selection coefficient and takes a simple form, with

$$\mathrm{G}(v) = \exp(-2v) \qquad (2.13)$$

(Figure 2.3A; Appendix 2, Section 3.2). In contrast, in the effectively neutral range ($S < 1$),

$$\mathrm{G}(v) = \sqrt{1 - v/v_{max}}, \qquad (2.14)$$

where the dependency on the selection coefficient enters through $v_{max} = \frac{1}{8}S$, which is the maximal contribution to variance and corresponds to an allele frequency of ½ (Figure A2.4A; Appendix 2, Section 3.2). In the intermediate selection regime, $\mathrm{G}(v)$ is also intermediate and

takes a more elaborate functional form (Appendix 2, Section 3.2). These results suggest how genetic variance would be distributed among sites given any distribution of selection coefficients (Figure 2.3A): starting from sites that contribute the most, the distribution would at first be dominated by strongly selected sites, then the intermediate selected sites would begin to contribute, whereas effectively neutral sites would enter only for $v < \frac{1}{8}S \ll 1$.

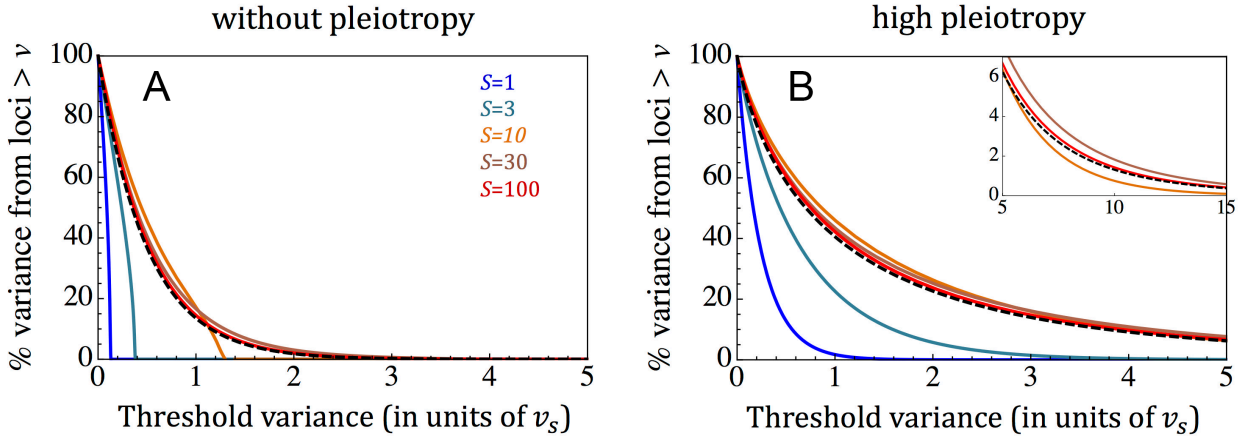



**Figure 2.3. The proportion of additive genetic variance that arises from sites that contribute more than the value on the x-axis, for a single trait (A) and in the pleiotropic limit (B).** Our approximations for sites under strong selection (Eqs. 2.12 & 2.14) are shown with the dashed black curves. For the approximations in the effectively neutral limit (Eqs. 2.14 and 2.16), see Figure A2.4.

Pleiotropy causes sites with a given selection coefficient to have a distribution of effect sizes on the focal trait, thereby increasing the contribution to genetic variance of some sites and decreasing it for others. In Appendix 2 (Section 3.2), we show that increasing the degree of pleiotropy, $n$, increases the proportion of genetic variance, $\mathrm{G}(v)$, for any threshold, $v$, regardless of the distribution of selection coefficients (Figure A2.5). When variation in a trait is sufficiently pleiotropic for the distribution of effect sizes to attain the limit form (Eq. 2.11):

$$\mathrm{G}(v) = \left(1 + 2\sqrt{v}\right)\exp\left(-2\sqrt{v}\right) \qquad (2.15)$$

for strongly selected sites and

$$\mathrm{G}(v) = \exp(-4\,v/S) \qquad (2.16)$$

for effectively neutral ones (Figures 2.3B and A2.4B; Appendix 2, Section 3.2). The intermediate selection range is split between these behaviors: on the weaker end, roughly corresponding to $S < 5$, $\mathrm{G}(v)$ is similar to the effectively neutral case (Figure A2.4B and Section 3.2 in Appendix 2); and on the stronger end, roughly corresponding to $S > 5$, $\mathrm{G}(v)$ is similar to the case of strong selection, with measurable differences only when $v \gg v_s$ (inset in Figure 2.3B and Section 3.2 in Appendix 2). We would therefore expect that as the sample size of GWAS increases and the threshold contribution to variance decreases, intermediate and strongly selected sites (more precisely sites with $S > 5$) will be discovered first, and effectively neutral sites will be discovered much later. In Appendix 2 (Section 3.2 and Figure A2.3), we also derive corollaries for the distribution of numbers of segregating sites that make a given contribution to genetic variance.

## Discussion

**Interpreting the results of human GWAS.** In humans, GWAS for many traits display a similar behavior: when sample sizes are small, the studies discover almost nothing, but once they exceed a threshold sample size, both the number of associations discovered and the heritability explained begin to increase rapidly [4,77]. Intriguingly though, both the threshold study size and rate of increase vary among traits. These observations raise several questions, including: How is the threshold study size determined? How should the number of associations and explained heritability increase with study size once this threshold is exceeded? With an order of magnitude increase in study sizes into the millions imminent, how much more of the genetic variance in

traits should we expect to explain? The theory that we developed provides tentative answers to these questions.

To relate the theory to GWAS, we must first account for the power to detect loci that contribute to quantitative genetic variation. In studies of continuous traits, the power can be approximated by a step function, where loci that contribute more than a threshold value $v^*$ to additive genetic variance will be detected and those that contribute less will not (Appendix 2, Section 6.1). The threshold depends on the study size, $m$, and on the total phenotypic variance in the trait, $V_P$, where $v^* \propto V_P/m$ [Appendix 2, Section 6.1; 77] conversely, the study size $m$ needed to detect loci with contributions above $v^*$ is proportional to $V_P/v^*$. Given a trait and study size, the number of associations discovered and heritability explained then follow from our predictions for the distribution of variances among sites.

When genetic variation in a trait is sufficiently pleiotropic, our results suggest that the first loci to be discovered in GWAS will be intermediate or strongly selected, with correspondingly large effect sizes (i.e., $S \approx \frac{2Nn}{w^2}a_1^2 > 5$). The functional form of the distribution of variances among these loci (Eq. 2.15 & Figure 2.3B) implies that for GWAS to capture a substantial proportion of the genetic variance, their threshold variance $v^*$ for detection has to be on the order of the expected variance contributed by strongly selected sites, $v_s$, or smaller. We therefore expect the threshold study size for the discovery of intermediate and strongly selected loci to be proportional to $V_P/v_s$. When the study size exceeds this threshold, the number of associations detected and proportion of variance explained depend on the study size measured in units of $V_P/v_s$ (Figure 2.4), and follow from the functional forms that we derived (Eq. 2.15 and Table A2.1). The dependence on $V_P/v_s$ makes intuitive sense, as the total phenotypic variance $V_P$ is background noise for the discovery of individual loci whose contributions to variance are on the

order of $v_s$. Some results are modified when variation in a trait is only weakly pleiotropic, which is probably less common: notably, the threshold study size for strongly selected loci would be higher and loci under intermediate selection would begin to be discovered only after the strongly selected ones (Figure A2.22 and Eqs. 2.13 and A2.35). Regardless of the degree of pleiotropy, effectively neutral loci would only begin to be discovered at much larger study sizes, after the bulk of intermediate and strongly selected variance has been mapped (Figures 2.4 and A2.22). Thus, the dependence of the explained heritability on study size is largely determined by $V_P/v_s$ and by the proportion of heritable variance arising from intermediate and strongly selected loci, whereas the number of associations also depends on the mutational target size, providing a tentative explanation for why the performance of GWAS varies among traits.

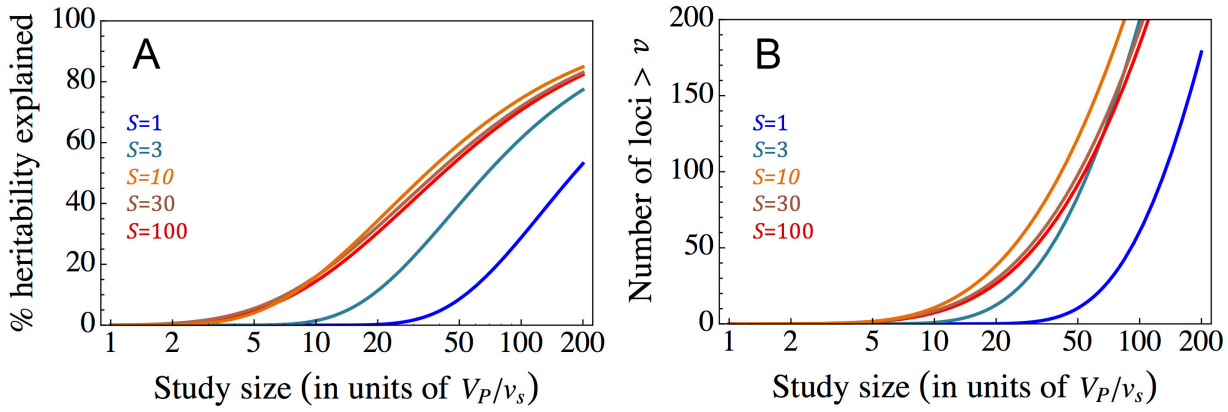



**Figure 2.4**. **The proportion of heritability (A) and the number of variants (B), assuming a mutational target size of 1 Mb, identified in GWAS as a function of study size,** in the pleiotropic limit (see Appendix 2, Sections 3.3 and 6.1, for derivations). For the case without pleiotropy, see Figure A2.22.

**Inference and prediction.** Importantly, these theoretical predictions can be tested. As an illustration, we consider height and body mass index (BMI) in Europeans, two traits for which GWAS have discovered a sufficiently large number of genome-wide significant (GWS) associations (697 for height [16] and 97 for BMI [15]) for our test to be well powered. We fit our

theoretical predictions to the distributions of variances among GWS associations reported for each of these traits, assuming that these distributions faithfully reflect what they would look like for the true causal loci (see Appendix 2, Section 6.3). We further assume that these loci are under intermediate or strong selection (as our predictions suggest) and that they are highly pleiotropic see Introduction; [37,42,45]. Under these assumptions, we expect the distribution of variances to be well approximated by a simple form (Eq. A2.89), which depends on a single parameter, $v_s$. We find that the theoretical distribution with the estimated $v_s$ fits the data for both traits well (Figure 2.5A): we cannot reject our model based on the data for either trait (by a Kolmogorov-Smirnov test, $p = 0.14$ for height and $p = 0.54$ for BMI; Appendix 2, Section 7.5). By comparison, without pleiotropy ($n$=1), our predictions provide a poor fit to these data (by a Kolmogorov-Smirnov test, $p < 10^{-5}$ for height and $p = 0.05$ for BMI; Figure A2.14).

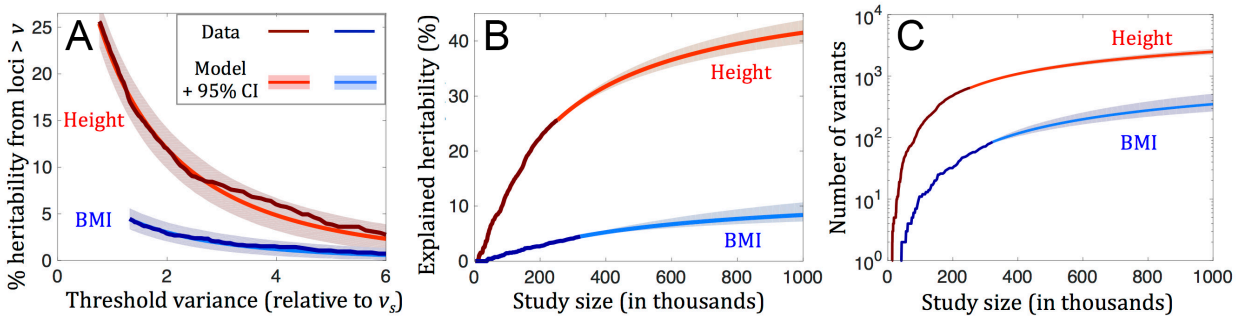



**Figure 2.5. Model fit and predictions for height and BMI, based on data from [16] and [15], respectively.** In (A), we show the fit for associated loci. In (B) and (C), we show our predictions for future increases in the heritability explained and number of variants identified as GWAS size increases. 95% CIs are based on bootstrap; see Appendix 2 (Section 7.4) for details.

Fitting the model to GWAS results further allows us to make inferences about evolutionary parameters (Appendix 2, Sections 7.1 and 7.3). By including the degree of pleiotropy ($n$) as an additional parameter, we find that for both height and BMI, $n$ is sufficiently large for it to be indistinguishable from the high pleiotropy limit. Based on the shape of the distributions in this

limit and on scaling the threshold values of $v^*$ in units of our estimates for $v_s$, we estimate that the proportion of variance arising from mutations within the range of detectable selection effects is ~50% for height and ~15% for BMI. Further relying on the number of associations that fall above the thresholds, we infer that, within this range, height has a mutational target size of ~5 Mb, whereas BMI has a target size of ~1 Mb (Table A2.2).

These parameter estimates can help to interpret GWAS results. They suggest that, despite their comparable sample sizes, the GWAS for height succeeded in mapping a substantially greater proportion of the heritable variance than the GWAS for BMI (~20% compared to ~3-5%) primarily because the proportion of variance arising from mutations within the range of detectable selection effects for height is much greater than for BMI. Moreover, the estimates of target sizes and the relationship between sample size and threshold contribution to variance can be used to predict how the explained heritability and number of associations should increase with sample size (Figure 2.5B-C). These predictions are likely under-estimates as the range of detectable selection effects itself should also increase with sample size.

We can also examine to what extent our inferences are consistent with data and estimates from earlier studies. For example, the distribution of variances that we inferred for height fits those obtained in a recent GWAS of height based on exome genotyping (Kolmogorov-Smirnov test, $p = 0.99$; Figure A2.15B and Section 8.1 in Appendix 2). In addition, the proportion of variance that we estimate to arise from the range of selection effects detectable in existing GWAS for height and BMI are consistent with estimates of the heritable variance tagged by all SNPs with MAF>1% [60,61; Appendix 2, Section 8.2].

**The effect of polygenic adaptation.** While we have assumed that quantitative traits have been subject to long-term stabilizing selection, recent studies indicate that some traits, and height in particular, have also been subject to recent directional selection [78-82]. Under plausible evolutionary scenarios, recent directional selection can induce large changes to the mean phenotype through the collective response at many segregating loci, while having a negligible effect on allele frequencies at individual loci [21,83]. This very subtle effect on allele frequencies is likely one reason why polygenic adaptation is so difficult to detect, and why studies have to pool faint signals across many loci to do so [78-82]. In Appendix 2 (Section 5.1), we show that the distribution of allele frequencies on which our results rely is insensitive to sizable recent changes to the optimal phenotype. Importantly then, even when recent directional selection has occurred and its effects are discernable, the genetic architecture of a trait is nonetheless likely to be dominated by the effects of longer-term stabilizing selection.

**The effect of demography.** In contrast, recent changes in the effective population size are likely to have had a dramatic effect on allele frequencies and thus on the genetic architecture of quantitative traits [84,85]. In particular, European populations in which the GWAS for height and BMI were performed are known to have experienced dramatic changes in population size, including an Out of Africa (OoA) bottleneck ~100 KYA and explosive growth over the past ~5 KY [86-89]. To study how these changes would have affected genetic architecture, we simulated allelic trajectories under our model and historical changes in population sizes in Europeans (relying on the model of [89]; Appendix 2, Section 9).

Our results suggest that the individual segregating sites with the greatest contribution to the extant genetic variance have selection coefficients around $s = 10^{-3}$ and are due to mutations

that originated shortly before or during the OoA bottleneck (Figure 2.6A and Section 9 in Appendix 2). These mutations ascended to relatively high frequencies during the bottleneck and minimally decreased in frequency during subsequent, recent increases in population size, thereby resulting in large contributions to current genetic variance. Segregating sites under weaker selection contribute much less to variance because of their smaller effect sizes (i.e., for the same reason that applied in the case with a constant population size). Finally, and in contrast to the case with a constant population size, individual segregating sites under stronger selection (e.g., $s \geq 10^{-2.5}$) contribute much less to current variance than those with $s \approx 10^{-3}$. Mutations at these sites are younger, and arose after the bottleneck, when the population size was considerably larger, resulting in much lower initial and current frequencies, and therefore a lower per (segregating) site contribution to variance as distinct from the proportion of strongly selected sites that are currently segregating, which will have greatly increased, resulting in the same total contribution to variance; [84,85]. In Appendix 2 (Section 10), we discuss one implication of these demographic effects: that the reliance on genotyping rather than resequencing in GWAS had a minimal effect on the discovery of associations.

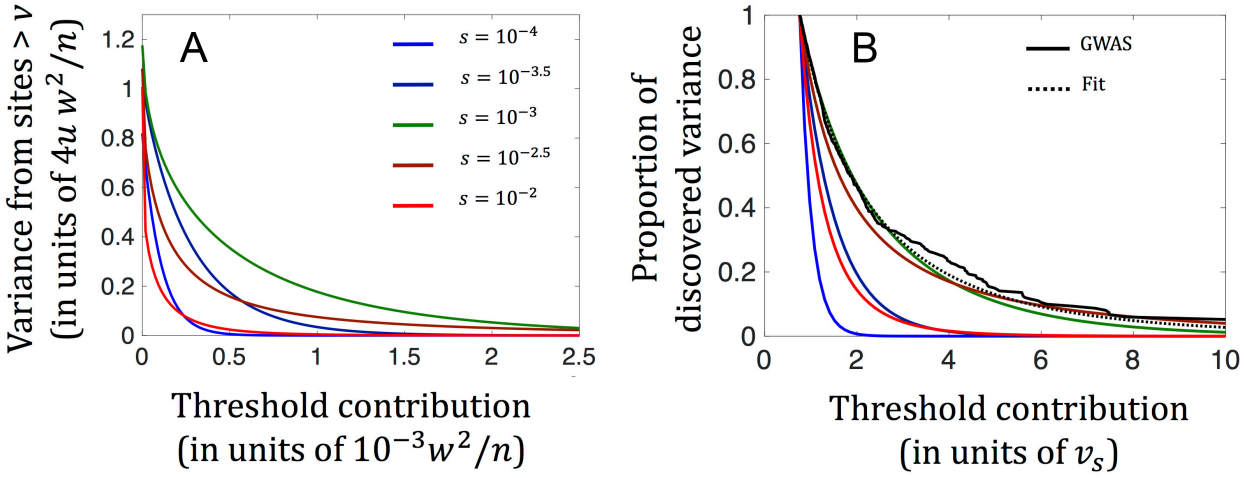


**Figure 2.6. The combined effect of selection and changes in population size (as inferred by [89] for Europeans) on the distribution of variances among segregating sites.** (A) The cumulative variance arising from sites with contributions above a threshold as a function of the threshold, for different selection coefficients. Cumulative variance is measured in units of $4u \cdot w^2/n$, the equilibrium expectation for a strongly selected site, while the threshold is in units of $10^{-3} \cdot w^2/n$. (B) The distribution of variances among loci identified in the GWAS of height. The empirical distribution is in solid black and our inferred fit is in dashed black. Simulation results for each selection coefficient (in color) are normalized such that the proportion of variance at the study threshold is always 1. For similar results corresponding to BMI, see Figure A2.20B, and for further details see Section 9 in Appendix 2.

Segregating loci with $s \approx 10^{-3}$ not only make the largest contributions to the current variance, but are also likely to account for most of the GWS associations in the GWAS of height and BMI (Appendix 2, Section 9). When we account for the discovery thresholds of these studies, the expected distribution of variances for loci with $s \approx 10^{-3}$ closely matches the distribution observed among GWS associations (Figures 2.6B & A2.20B). Moreover, these distributions closely match our theoretical predictions for $s \approx 10^{-3}$ and an $N_e \approx 5000$ (Figure 2.6B)—roughly the effective population size experienced by mutations that originated shortly before or during the bottleneck. This match likely explains why the results predicted on a constant population size fit the data well nonetheless. Our interpretation of GWAS findings is supported by other aspects of the data (Appendix 2, Section 9).

Our conclusions about the high degree of pleiotropy of genetic variation for height and BMI and the differences between these traits are likely robust to demographic effects, given how well our model fits the distributions of variances among loci, once we account for European demographic history. However, we might be underestimating the mutational target sizes and total heritable variances associated with the selection effects currently visible in GWAS, as simulations with European demographic history indicate that the proportion of variance arising from loci with $s \approx 10^{-3}$ explained by current GWAS is lower than our equilibrium estimates (~42% compared to

~53% for height, and ~29% compared to ~38% for BMI). By the same token, we likely underestimated the future increases in explained heritability with increases in study sizes (Figure 2.5B-C).

## Conclusion

In summary, a ground-up model of stabilizing selection and pleiotropy can go a long way toward explaining the findings emerging from GWAS. Important next steps involve explicitly using more information from GWAS in the inferences. In particular, we can learn more about the selection acting on quantitative genetic variation by explicitly incorporating information about frequency and effect size (rather than their combination in terms of variance), and by including information from associations that do not attain genome-wide significance. Doing so will further require directly incorporating the effects of recent demographic history on genetic architecture [84,85]. An extended version of the inference, applied to the myriad traits now subject to GWAS, should allow us to learn about differences in the genetic architectures of traits, and answer long-standing questions about the evolutionary forces that shape quantitative genetic variation.

## Acknowledgements

We have benefited hugely from discussions and comments from Guy Amster, Nick Barton, Jeremy Berg, Graham Coop, Laura Hayward, David Murphy, Joe Pickrell, Jonathan Pritchard and Molly Przeworski.

# Chapter 3
# Inferring Selection on Human Quantitative Genetic Variation


## Abstract

**Many human traits exhibit considerable heritable variation in natural populations, generated by numerous genetic variants with small effects. We know very little about the selection acting on such variants, mainly because only recently have genome-wide association studies (GWAS) begun uncovering the genetic basis of quantitative traits. We have argued in our previous work that quantitative traits are often under pleiotropic stabilizing selection and have developed equations to describe the effect of natural selection on trait-affecting alleles. Here, we extend our work to infer the distribution of selection coefficients acting on variants discovered in GWAS. Our method uses both the frequencies and effect sizes of genome-wide significant GWAS associations ("GWAS hits"), which gives it considerable statistical power. We investigate the performance of the inference on simulated datasets, aimed at mimicking human GWAS data under a variety of assumptions about demography and selection. The method does extremely well at estimating both the mean and standard deviation of selection coefficients and even does well at inferring the distribution of selection coefficients. We translate our estimates of the distribution of selection coefficients for GWAS hits into the distribution of selection coefficients for newly arising mutations within a range of selection coefficients in which GWAS is well-powered to identify variants. For this range, we predict the increases in explained heritability and number of associations as GWAS sample sizes increase. This is**

**the first method to infer the distribution of selection effects acting on variants contributing to quantitative trait variation. We are currently working on applying our method to GWASs in the UKBiobank for human quantitative traits.**

## Introduction

Many traits that we care about exhibit considerable heritable variation in natural populations, which arises from numerous segregating alleles of small effect[1]. In particular, quantitative traits, including height and BMI, are highly heritable and genetically complex. Despite over a century of theoretical and empirical research, we still know little about the population genetic processes that give rise to variation in such traits[2]. In particular, we know little about the selection acting on variants contributing to quantitative trait variation[3].

Perhaps the main obstacle is that until recently we knew little about the genetic basis of quantitative traits[4]. Over the last decade, genome-wide association studies (GWAS) in humans (and other organisms) have begun to systematically map the genetic variants underlying quantitative trait variation. In humans, GWAS have already made tens of thousands of reproducible associations between genetic variants (mostly SNPs) and quantitative traits.

The associations made by GWAS provide an unprecedented opportunity to learn about the evolutionary forces that shape quantitative trait variation[5]. GWAS are uncovering the genetic architecture of traits, i.e. the number, frequency, effect size and distribution of variants underlying trait variation. Genetic architectures have been shaped by evolutionary forces, chiefly by mutation, selection and genetic drift. Therefore, the information GWAS are providing about genetic architectures can be used to learn about these evolutionary processes.

Specifically, the frequency and effect sizes of GWAS associations reflect the selection acting on a trait and their number reflects the mutational input. Selection (together with genetic drift) determines the distribution of allele frequencies[6]. The relationship between the effect of an allele on a trait and the selection acting upon it depends on the nature of selection. For example, if a trait is always selected for or always selected against (directional selection), selection coefficients scale linearly with effect size. Conversely, if selection acts to maintain an optimal trait value (stabilizing selection), selection coefficients scale like the effect size squared. In turn, the number of variants seen across the genome is a direct consequence of the number of potential trait-affecting mutations, i.e. the "mutational target size".

In our previous work[7], we argued that quantitative traits are often under pleiotropic stabilizing selection and developed equations to describe the effect of selection on trait-affecting alleles. These equations led to simple and robust predictions for the contributions to variance of the trait-associated variants seen in GWAS. These predictions proved to be a near perfect fit to the findings of GWAS for height and BMI, using a single free parameter. This fit came as somewhat of a surprise since our predictions were based on equilibrium demography. We argued, via simulations, that this good fit is the product of the narrow range of selection coefficients that should be accessible in GWAS. Here, we extend our work to infer the distribution of selection coefficients seen in GWAS.

The idea of inferring the distribution of selection effects from information about segregating genetic variation has been explored extensively in population genetics over the past two decades. Sawyer & Hartl, in their seminal work[8], suggested using the distribution of frequencies (the site frequency spectrum) of non-synonymous mutations in genes to infer the distribution of selection effects of new mutation in genes. Once genomic datasets started becoming available, many

works implemented Sawyer & Hartl's ideas and attempted to estimate the distribution of selection coefficients at newly arising mutations in a variety of species [9-13].

Our method is distinct from these methods in its objective and in the data it uses. We seek to infer the distribution of selection coefficients not of all mutations but only of mutations that affect a given trait. The number of such mutations reliably identified by GWAS, usually hundreds or thousands, is smaller by orders of magnitude than the number of segregating mutations used by the other methods. However, we have the advantage of using not only frequencies but also effect sizes in our inference, and this combined information is what gives our method its power.

## The inference method

Our inference uses the frequencies and effect sizes of genome-wide significant GWAS associations ("GWAS hits") to build a composite likelihood of the distribution of selection effects. If we knew the frequencies and effect sizes of all trait affecting variants, we could write down the following composite likelihood function for the distribution of selection coefficients:

$$\mathrm{L}(\mathrm{f}(s)|\{x_i, a_i\}) = \prod_{\mathrm{i}} \int_s \mathrm{P}(\mathrm{x_i}|\mathrm{s})\mathrm{P}_c(\mathrm{a}_i|\mathrm{s})\mathrm{f}(s) \qquad (3.1)$$

with $\mathrm{P}(\mathrm{x_i}|\mathrm{s})$ being the frequency distribution (SFS) for selection coefficient $s$, $\mathrm{P}_c(\mathrm{a}_i|\mathrm{s})$ being the distribution of effect sizes for selection coefficient $s$, and $\mathrm{f}(s)$ being the distribution of selection coefficients that we are trying to estimate. $\mathrm{P}(\mathrm{x_i}|\mathrm{s})$ is given by our equations for change in allele frequencies under stabilizing selection[7] and can be calculated analytically for equilibrium demography or estimated via simulations or numerically for non-equilibrium demography. We assume $\mathrm{P}_c(\mathrm{a}_i|\mathrm{s})$ takes its high pleiotropy limit and it depends on a scaling parameter $c$ that measures effect size in units of selection. For more details, see the Methods Section & Appendix 3.

As can be seen in Figure 3.1, the functional form of this likelihood (Eq. 3.1) illustrates the power of using the co-distribution of frequencies and effect sizes. Selection limits the frequency an allele can reach and therefore seeing an allele at a specific frequency limits the selection coefficients that can plausibly be acting on it. Therefore, frequencies act as an upper bound on selection coefficients (Fig. 3.1A). On the other hand, effect sizes act as lower bounds on selection coefficients since selection is driven by the size of the effect an allele has on traits (see Fig. 3.1B). (Effect sizes also provide a weak, polynomial upper bound.) The result is that, in our model, the frequency and effect size of an allele, together, are enough to bound the selection coefficient acting on it (Fig 3.1C).

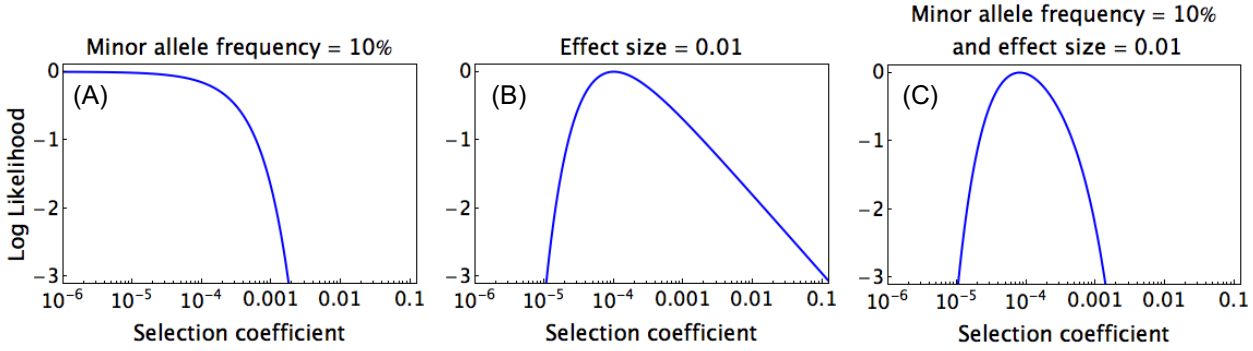



**Figure 3.1. The combined information of frequency and effect size provides considerable information about the selection coefficient.** The log likelihood of the selection coefficient conditional on a variant's (A) minor allele frequency (B) effect size and (C) both minor allele frequency and effect size. Shown for a constant population size of $N_e = 10{,}000$ and $c = 1$, see Eq. 3.1. Likelihood measured relative to its maximum.

The likelihood has to be adjusted to take into account the limited power of GWAS[14]. GWAS are only well-powered to discover variants contributing more than a threshold contribution, $v^*$, to variance, i.e., those with $v = 2a^2x(1-x) > v^*$. Moreover, since GWAS rely on genotyping they also having an effective cutoff for frequencies, i.e., $x > x^*$. Therefore, the composite likelihood conditional on discovery in GWAS is

$$\mathrm{L}(\mathrm{f}(s)|\{x_i, a_i\}) = \prod_{\mathrm{i}} \int_{\mathrm{s}} \mathrm{P}(\mathrm{x_i}|\mathrm{s})\mathrm{P_c}(\mathrm{a_i}|\mathrm{s})\mathrm{f}(s) \,/\, \int_s \mathrm{n_c}(v^*, \mathrm{s})\mathrm{f}(s) \qquad (3.2)$$

with

$$\mathrm{n_c}(v^*, \mathrm{s}) = \int_{x,a} \mathrm{P(x|s)P}_c(\mathrm{a|s})\theta(\mathrm{x} - \mathrm{x}^*)\theta(v - v^*) \quad (3.3)$$

being the proportion of sites under selection coefficient $s$ that is discovered in GWAS and θ being the Heaviside step function.

Inferring $\mathrm{f}(s)$ is impractical since GWAS only discover variants from a limited range of selection coefficients. For very small selection coefficients, effect sizes will be too small to be identified in GWAS, while for large selection coefficients, alleles will be too rare. Therefore, we expect to have no data for these ranges of selection coefficients and trying to infer $\mathrm{f}(s)$ at these ranges is bound to cause numerical and conceptual problems.

We therefore choose instead to estimate the distribution of selection effects of GWAS hits, $\mathrm{g}(s)$. We reparametrize the likelihood in terms of $\mathrm{g}(s)$:

$$\mathrm{L}(\mathrm{g}(s)|\{x_i, a_i\}) = \prod_{\mathrm{i}} \int_s \frac{\mathrm{P(x_i|s)P}_c(\mathrm{a_i|s})}{\mathrm{n_c}(v^*,s)} \mathrm{g}(s) = \prod_{\mathrm{i}} \int_s \mathrm{P}_c(x_i, a_i|v^*, s)\mathrm{g}(s) \quad (3.4)$$

with $\mathrm{g}(s) = \mathrm{f}(s)n_c(v^*, s)/\int_s \mathrm{f}(s)n_c(v^*, s)$ and $\mathrm{P}_c(\mathrm{x}, \mathrm{a}|v^*, s) = \frac{\mathrm{P(x_i|s)P}_c(\mathrm{a_i|s})}{\mathrm{n_c}(v^*,\mathrm{s})}$ being the distribution of frequencies and effect sizes conditional on selection coefficients and discovery in GWAS.

From this point on, inferring the distribution of selection coefficients of GWAS hits is a purely technical challenge. To keep the shape of the distribution $\mathrm{g}(s)$ as simple as possible, we parameterize the distributions in terms of the $\log_{10}$ of selection coefficients. However, the maximum-likelihood estimator tends to be very spiky, so we use a log spline smoothing method[15] to make sure the inferred $\mathrm{g}(s)$ is smooth and well-behaved. For equilibrium demography, the likelihood can be calculated analytically, whereas for non-equilibrium

demography, $P(x|s)$ cannot be calculated analytically; we therefore use extensive simulations to estimate it. We maximize the likelihood using the Nelder-Mead algorithm[16] but since it is impractical to integrate over $s$ in Eq. 3.4 during the maximization process, we approximate the integral by a Riemann sum over a grid of selection coefficients. For more details, see the Methods Section & Appendix 3.

## Results

We investigated the performance of the inference on simulated datasets, aimed at mimicking human GWAS data under a variety of assumptions about demography and selection (see Methods and Appendix 3). We considered five different distributions of selection coefficients at newly arising trait affecting mutations: three single Gamma distributions with shape parameter 1 (exponential distributions), which on a $\log_{10} s$ scale are unimodal; one more complex mixture of three Gamma distributions, which is still unimodal on a $\log_{10} s$ scale; and one bimodal mixture of three Gamma distributions. For each of these distributions, we considered two demographic scenarios: a constant population size of $N_e = 10{,}000$ and a demographic model with changing population sizes previously estimated for European populations [7,17]. To simulate the set of effect sizes and frequencies corresponding to GWAS hits, we choose several values of the scaling parameter, c, and set the threshold contribution to variance above which variants are discovered to $v^* = 10^{-4}$ in units of trait phenotypic variance, corresponding to a study size of approximately 300,000. Together, $v^*$ and $c$ determine the expected variance captured by GWAS, allowing us to examine how our inference depends on the heritability explained by GWAS hits. We then sampled selection coefficients from the distribution of selection effects, assigning them frequencies and effect sizes according to $P(x|s)$ and $P_c(a|s)$, respectively (Eqs. A3.3 & A3.4). We include a simulated variant in our set only if it meets the conditions $v = 2a^2x(1-x) > v^*$

and $x > x^*$. We continued the sampling until we obtain 2,000 GWAS hits, roughly corresponding to the number of hits we expect to see for height in the UK Biobank[18], based on our previous predictions. For each combination of parameters (i.e., distribution of selection effects, demographic model and scaling parameter), we performed 100 simulations.

The performance of our inference on simulated data are summarized in Figs. 3.2 & 3.3. The method does extremely well at estimating both the mean and standard deviation of selection coefficients under the range of scenarios considered (Fig. 3.2). Moreover, it performs well at inferring the distribution of selection effects (Fig. 3.3). While sometimes details of the probability density are missed (e.g., Fig. 3.3.B), the CDF of the distribution is captured very well nonetheless.

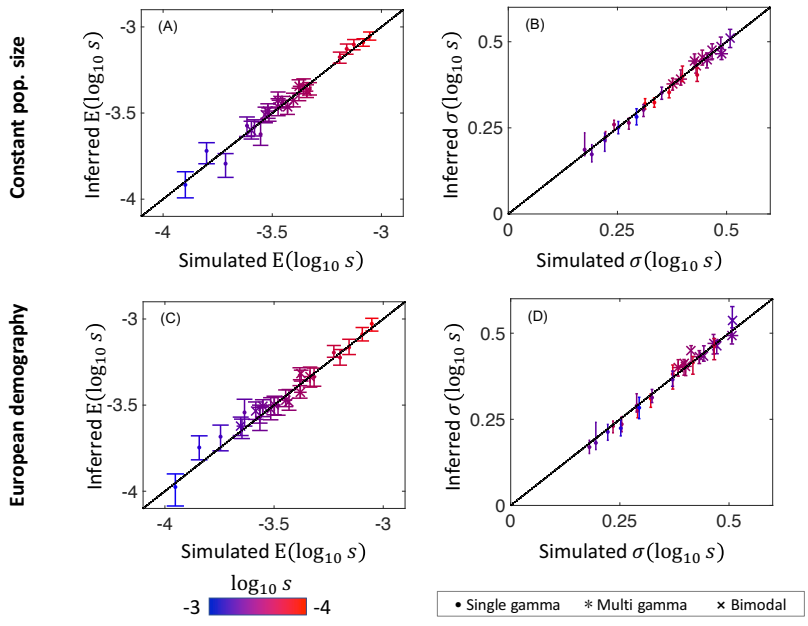



**Figure 3.2. The inferred mean and standard deviation of selection effects are both accurate and precise.** Selection coefficients are shown on a log scale. Comparison of the estimates of the mean (A&C) and standard deviation (B&D) of logged selection coefficients with the true parameters under a variety of scenarios (see text and Appendix 3 for details). Simulated with a constant population size of 10,000 (A-B) or European demography (C-D). Each point corresponds to a single simulation, where the widgets corresponds to 90% CI estimated over 100 bootstraps.

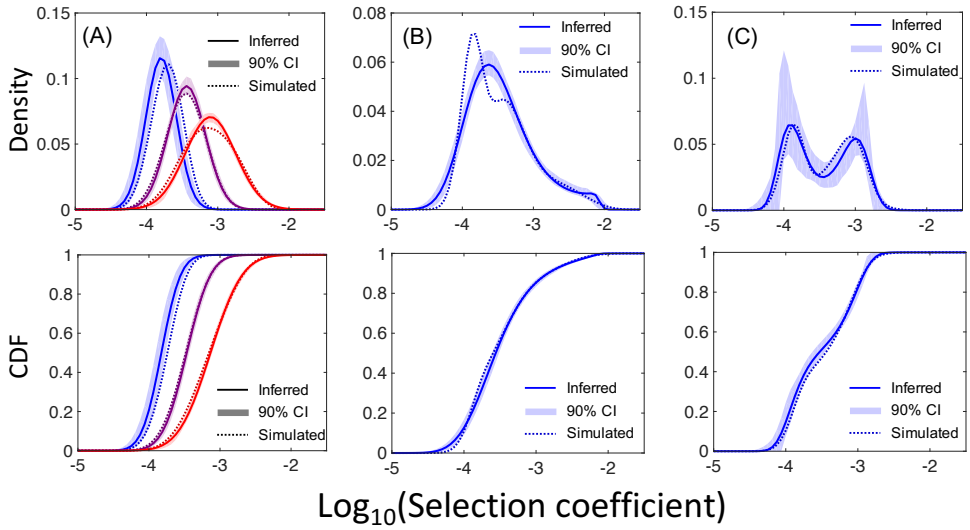



**Figure 3.3. The inference recovers the true distribution of selection coefficients.** The true, simulated distributions (doted) and inferred distributions (solid+90% CI sleeve) are shown for three kinds of distributions: single Gamma distributions, with shape parameter 1 (A), a unimodal mixture of three Gamma distributions (B), and a bimodal mixture of three Gamma distributions (C). In all of these simulations the other parameters where $c = 0.5$ and $v^* = 10^{-4}$. Shown here for a constant population size. See Appendix 3 for more details and equivalent figure under non-equilibrium demography.

We translated our estimates of the distribution of selection coefficients for GWAS hits into the distribution for newly arising mutations (Figure 3.4A). To this end, we calculated the probability that an allele with a given selection coefficient exceeds the power thresholds and will be discovered in a given GWAS (Eq. 3.3). The distribution of selection coefficients for newly arising mutations can then be calculated by dividing the estimated distribution for the observed variants, $g(s)$, by this probability of discovery (Fig. 3.4A).

Our inference only provides reliable estimates of the distribution of selection coefficients within a range in which GWAS is well-powered to identify variants (see [7]). For a population of constant size, this range is bounded from below since the contribution to variance of alleles with small selection coefficients falls under the threshold of GWAS and from above since large selection coefficients lead to low frequencies which can't be imputed. The upper bound becomes more pronounced under non-equilibrium demography, since recent explosive population growth shifts

strongly selected alleles to exceedingly low frequencies, driving them below both the frequency and contribution to variance thresholds. Outside this range, the probability of detecting a variant in GWAS becomes incredibly small. We divide by this probability in order to get the distribution of selection coefficients for newly arising mutations. Therefore, for the ranges of selection coefficients in which the GWAS is underpowered the confidence intervals for our estimates become extremely large (Fig, 3.4B). For these reasons, we define the range in which our estimates are reliable as the range which is predicted to contain all but 100 of the GWAS hits, ensuring that within this observed range GWAS is well-powered to detect variants.

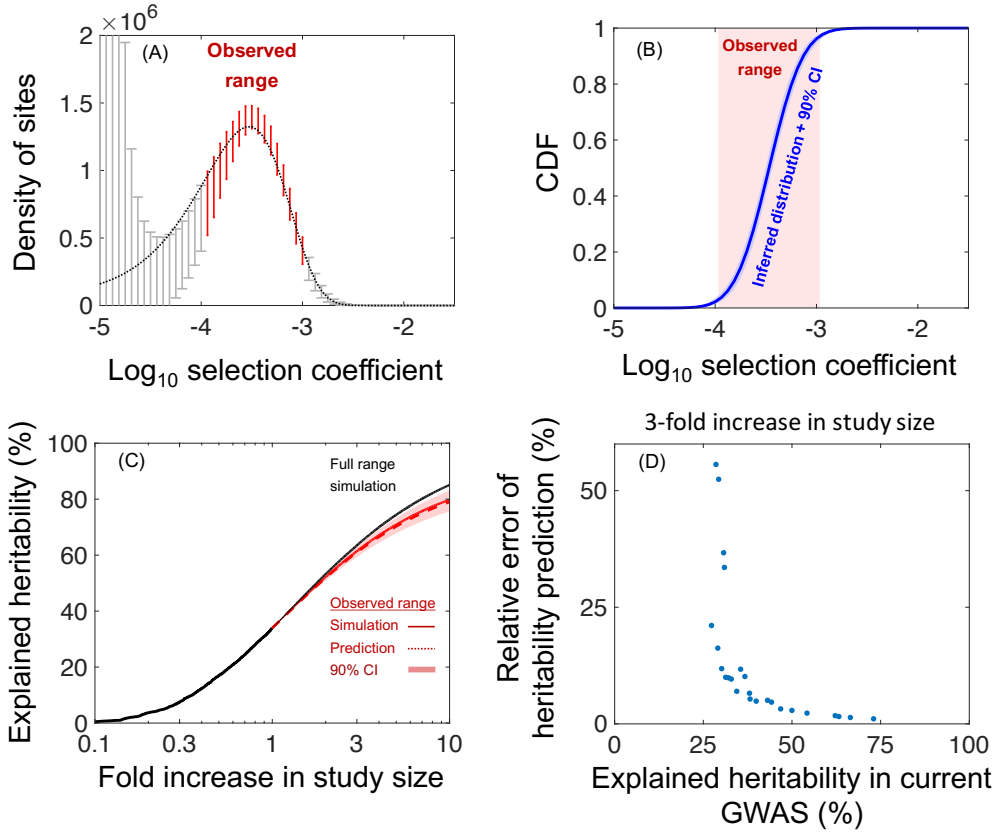



**Figure 3.4. Predictions based on the inferred distributions.** (A) The inferred distribution of selection coefficients of newly arising mutations. The confidence intervals for our estimates become very large outside the range that GWAS is well-powered to detect. (B) Comparison of the CDF of the distribution of selection coefficient within the range in which the method is well-powered. See text for how we define this range. (C) The increase in explained heritability as a function of the increase in study size. The prediction and contribution from the observed range are in red and simulation results for the full range are in black. (D) Relative error of heritability

prediction (ratio of 90% CI to predicted heritability) for a three-fold increase in study size as a function of explained heritability in current GWAS. (A-C) are for $c = 0.5$ with an exponential distribution of selection coefficients for newly arising mutation with mean $E(s) = 3 \cdot 10^{-4}$. All results are with $v^* = 10^{-4}$ and a constant population size of $N_e = 10^4$. See Appendix 3 for details and equivalent figure under non-equilibrium demography.

We used our estimate of the distribution of selection coefficients to predict the increases in explained heritability and number of associations as GWAS sample sizes increase (Figure 3.4C, Figure A3.4 & Figure A3.5). For modest increases in study size, these projections are extremely accurate but as study sizes increase the uncertainty in our projections also increases, reflecting the uncertainty in our estimation of the distribution of selection coefficients and the scaling parameter. The accuracy of our predictions crucially depends on the proportion of variance currently explained by GWAS, with well-powered GWAS producing much more accurate predictions (Figure 3.4D). As study sizes increase, the observable range of selection coefficients also increases but our projections consider only contributions from the original, more limited range and therefore tend to underestimate the number of variants and heritability that will be observed.

## Discussion

To the best of our knowledge, this method to infer the distribution of selection effects acting on variants contributing to quantitative trait variation based on GWAS data is the first of its kind. We illustrated that it should work with the kind of data now available from GWAS of many traits, and qualified its limitations, notably in terms of the range of selection effects that it can infer. We also illustrated how the inferences can be used to predict the findings of future GWAS studies.

Our inference methodology resolves several crucial problems in previous methods used to infer distributions of selection effects[19]. Our method is almost completely non-parametric, as the only assumptions it makes on the distribution of selection effects is smoothness. In addition, our method clearly delineates from which range of selection coefficient our signal arises and for which selection coefficients our inference is informative. These features protect our results from artifacts and overinterpretation.

Our inferences rely on the generative model presented in Simons et al[7]. The predictions of this model are robust to many factors, including moderate asymmetry in mutational input, recent changes in trait optimum, the existence of major effect loci, and most forms of anisotropic mutation (see Simons et al. [7], SI Section 5). Moreover, when these assumptions are violated, it should be straightforward to apply our inference approach to modified or alternative generative models that describe the distributions of frequencies and effect sizes conditional on the selection coefficient $\left(\mathrm{P}(\mathrm{x}|\mathrm{s}) \text{ and } \mathrm{P}(\mathrm{a}|\mathrm{s})\right)$.

Our statistical method can also be easily extended to deal with some sources of error in GWAS. While allele frequency estimates in GWAS are likely to be highly precise because of the enormous sample sizes, estimates of effect sizes are known to suffer from considerable uncertainty[14]. These errors are easy to account for in our method by replacing $\mathrm{P}(\mathrm{a}|\mathrm{s})$, the probability of a variant having effect size $a$ conditional on having selection coefficient $s$, with $\mathrm{P}(\hat{\mathrm{a}}|\mathrm{s})$, the probability of a GWAS estimating a variant as having effect size $\hat{a}$ conditional on having selection coefficient $s$. Such probabilities are easy to calculate since GWAS sampling errors are approximately normally distributed (see Appendix 3, Section 10).

Perhaps the greatest concern in applying our method to real data is the discrepancy between the effect sizes and frequencies of association identified in GWAS and the causal variants that they

tag. Our method assumes that they are one and the same. We considered applying current fine mapping algorithms[20-22] to estimate weights on possible sets of causal loci under GWAS peaks, and apply our inference to these sets. Unfortunately, it appears that most current methods for fine mapping are not scalable to more than a few peaks, and are ill-suited to our problem since their output, notably the number of potentially causal variants, strongly depends on tuning parameters. While we are still considering other approaches, it is possible that we will have no good way to characterize the uncertainty in frequency and effect size due to this problem.

While we may not be able to assess the uncertainty due to imperfect tagging, we may be able to test the veracity of our inferences against largely independent aspects of the data. For example, given our inferred distribution of selection coefficients and the frequencies of effect size of GWS associations, we can calculate the expected distribution of allelic ages. As new methods[23] have been recently developed to estimate the ages of SNPs based on haplotype data, we can compare the estimated distributions with those that we predict. As another closely related verification method, we can use our results to calculate, for variants captured by GWAS in one population, the expected frequency distribution in other populations, which can easily be compared to publicly available data.

We are currently working on applying our method to GWASs in the UKBiobank[18] for traits, including height, BMI, age at menarche, male pattern baldness and educational attainment. We hope to uncover the differences in selection effects and mutational target sizes underlying quantitative genetic variation in these traits and to understand the evolutionary drivers of these differences. Beyond its inherent interest, understanding the evolutionary determinants of the genetic architecture of traits can help design better association studies in the future and be used

as a prior for making clinically relevant predictions from current studies. We expect our results will prove meaningful from both a basic science and practical perspectives.

## Methods

**The composite likelihood.** Our inference is based on maximizing the composite likelihood of the distribution of selection effects given the frequency and effect size of variants identified in GWAS. Here, we describe how we construct the composite likelihood function. First, we define the probability distributions of variant frequencies and effect sizes for a given selection coefficient. We use them to produce a composite likelihood for the distribution of selection coefficients at newly arising mutations, using all trait-affecting variants as data. We then condition on variants being discovered by GWAS by renormalizing the likelihood. Finally, we reparametrize this likelihood in terms of the distribution of selection coefficients of GWAS hits. For a given selection coefficient, the distributions of variant frequencies and effect sizes are independent of each other and can be described by our model of stabilizing selection[7]. Our equation for change in allele frequency, coupled with a demographic model, translates to the distribution variant frequency $x$, $\mathrm{P}(x|s)$, for a given selection coefficient. For equilibrium demography, this distribution can be approximated by

$$\mathrm{P}(x|s) = 2Nu \cdot \tau(x|s) \qquad (3.5)$$

with $2Nu$ being the mutational input and $\tau$ being the sojourn time as calculated by the diffusion equation (see Appendix 2 for an analytic expression for $\tau$). For non-equilibrium demography, we use simulations to approximate $\mathrm{P}(x|s)$. We choose to use a folded frequency spectrum throughout, i.e., $x$ represents the minor allele frequency, since allows us to avoid issues of

ancestral allele misidentification and does not significantly affect our power. The density of effect size is

$$\mathrm{P}_c(a|s) = \frac{1}{\sqrt{2\pi c \cdot s}} \exp\left(-\frac{1}{2}\frac{a^2}{c \cdot s}\right) \quad (3.6)$$

with $c$ being a scaling constant. As shown in Simons et al. [7], under equilibrium demography $c$ is the expected contribution to variance from a strongly selected site. More generally, $c$ converts selection to units of effect size. Taken together, these two equation describe the distribution of frequencies and effect sizes $\mathrm{P}_c(x, a|s) = \mathrm{P}(x|s)\mathrm{P}_c(a|s)$.

Using the distribution of frequencies and effect sizes, we write down a composite likelihood for the distribution of selection coefficients at newly arising mutations, $\mathrm{f}(s)$, using trait-affecting variants as data:

$$\mathrm{L}(\mathrm{f}(s)|\{x_i, a_i\}) = \prod_{\mathrm{i}} \int_s \mathrm{P}_c(x, a|s)\mathrm{f}(s) \quad (3.7)$$

with $\{x_i, a_i\}$ being the frequencies and effect sizes of **all** trait affecting variants.

However, in practice we have access only to variants discovered by GWAS. As discussed above, we approximate the effect of discovery by GWAS as applying thresholds on the frequencies and contributions to variance of variants. Therefore, we renormalize the likelihood using an expression for the proportion of variants of a given selection coefficient discovered by a GWAS as a function of the threshold contribution to variance $v^*$ and threshold frequency $x^*$:

$$n_c(v^*, s) = \int_{v>v^*, x>x^*} \mathrm{P}_c(x, a|s). \quad (3.8)$$

Making the likelihood conditional on discovery in GWAS

$$\mathrm{L}(\mathrm{f}(s)|\{x_i, a_i\}) = \prod_{\mathrm{i}} \frac{\int_s \mathrm{P}_c(\mathrm{x_i}, \mathrm{a_i}|\mathrm{s})\mathrm{f}(s)}{\int_s \mathrm{n_c}(v^*, \mathrm{s})\mathrm{f}(s)} \quad (3.9)$$

with $\{x_i, a_i\}$ being the frequencies and effect sizes of **variants discovered by GWAS**.

We choose to infer the distribution of selection coefficients at GWAS hits instead of at newly arising mutations. We make this choice to avoid infering information about ranges of selection coefficients that are not well-covered by GWAS findings. We therefore define the distribution of frequencies and effect sizes from a given selection coefficient conditional on discovery in GWAS

$$\mathrm{P}_c(x, a|v^*, s) = \frac{\mathrm{P}_c(x,a|s)}{n_c(v^*,s)}\theta(v > v^*)\theta(x > x^*) \quad (3.10)$$

and we then rewrite the likelihood as

$$\mathrm{L}(\mathrm{g}(s)|\{x_i, a_i\}) = \prod_{\mathrm{i}} \int_s \mathrm{P}_c(\mathrm{x_i}, \mathrm{a_i}|v^*, s)\mathrm{g}(s) \quad (3.11)$$

with $\mathrm{g}(s) = \mathrm{f}(s)n_c(v^*, s)/\int_s \mathrm{f}(s)n_c(v^*, s)$ being the distribution of selection coefficients at GWAS hits. Note, that Eq. 3.11 takes the same form as Eq. 3.7, only with a conditional distribution of frequencies and effect sizes.

**Demography.** We use forward simulations to approximate the frequency distribution under non-equilibrium demography for each selection coefficient on our grid. The simulation procedure we use is described in Simons et al.[7]. For the results shown here, we assume the demographic model inferred by Schiffles and Durbin[17] for European populations. On our grid, we run 240 million simulations of biallelic sites, with a mutation rate of $u = 1.25 \cdot 10^{-8}$ per bp per generation. We use these simulations for two purposes: i) to calculate the $\mathrm{P}(x|s)$ term in the composite likelihood, where to this end, we bin allele frequencies (see details in Appendix 3) and ii) to produce simulatesd datasets (see below).

**Simulating GWAS data.** We simulate a given dataset as follows: 1) We draw a selection coefficient from our assumed distribution. 2) Given a selection coefficient on the grid, we choose

its effect size based on the distribution in Eq. 3.6 . For a constant population, we choose the frequency based on equation 3.5 whereas for European demography, we sample frequencies from our simulations. 3) We accept a variant if its contribution to variance is above $v^*$ and its frequency is above $x^*$. 4) We repeat this process until we have 2,000 variants.

We assume the distribution of selection coefficients for newly arising mutations is a sum of Gamma distributions, i.e., $\mathrm{f}(s) = \sum_j w_j \mathrm{p}_j(s|k_j, \theta_j)$, with $\mathrm{p}(s|k, \theta) = \frac{1}{\Gamma(k)\theta^k} e^{-s/\theta} s^{k-1}$ the standard Gamma distribution PDF and $w_j$ being weights. We choose the weights so GWAS hits have equal probability to come from each of the Gamma distributions. We use three single Gamma distributions, with $k = 1$ and $\theta = 10^{-4}, 3 \cdot 10^{-4}, 10^{-3}$. We also include two mixtures of three Gamma distributions: One mixture is unimodal on a $\log_{10} s$ scale and is a mixture of Gammas with $k = 5, 1, 0.3$ and $\theta = 2 \cdot 10^{-5}, 3 \cdot 10^{-4}$,$3.3 \cdot 10^{-3}$, respectively. The other mixture is bimodal on a $\log_{10} s$ scale and is a mixture of Gammas with $k = 5, 0.3, 5$ and $\theta = 2 \cdot 10^{-5}, 10^{-3}$,$2 \cdot 10^{-4}$, respectively.

**Maximizing the likelihood.** In order to maximize the likelihood, we have to employ a few simplifying assumptions. We replace the integral over $s$ in Eq. 3.4 by a Riemann sum over a dense grid of selection coefficients. We also assume that the distribution of contributions to variance from GWAS hits, $\mathrm{g}(s)$, is smooth on a $\log_{10} s$ scale. This is a reasonable assumption given that many GWAS hits come from a limited but continuous range of selection coefficients (see Appendix 3 for details).

In order to maximize the likelihood, we have to use a finite dimensional representation of $\mathrm{g}(s)$, so we parametrize $\mathrm{g}(s)$ as a logspline, i.e., represent its log by a cubic spline[15]. This representation has the advantage that it can represent very intricate functional forms, as long as the spline has enough knots. Knots are the breakpoints where the spline's third derivative can

change; a spline is completely defined by its value at those knots. To prevent spurious spikey features in $\mathrm{g}(s)$, we introduce a penalty in the log likelihood for each additional knot (see Appendix 3 for details).

Maximization of the likelihood proceeds in two stages. For each value of $c$ and set of spline knots, we use the Nelder-Mead algorithm[16] to maximize the likelihood. We use an MCMC method (similar to SALSA[24]) to sample values of $c$ and sets of spline knots and choose the one that produces the highest penalized likelihood.

**Predictions of future GWAS.** We can only make predictions from the range of selection coefficients observed in GWAS. After testing several ways to determine the range in which we are well-powered, we found that excluding all but 100 variants works well. More formally, this is the range at which the CDF of inferred distribution of selection coefficients at GWAS hits $\hat{\mathrm{g}}(s)$, $\widehat{\mathrm{G}}(s) = \int_{s'<s} \hat{\mathrm{g}}(s')$, is between $50/n$ and $1 - 50/n$. We define $s_{down}$ and $s_{up}$ as the lower and upper bound of this range, that is $\widehat{\mathrm{G}}(s_{down}) = 50/n$ and $\widehat{\mathrm{G}}(s_{up}) = 1 - 50/n$. This choice is designed to use all the selection coefficients for which we have information; using simulations, we have seen it roughly corresponds to including regions where the 90% CI of $\hat{\mathrm{g}}(s)$ is smaller than $\hat{\mathrm{g}}(s)$.

We can predict the number of associations made in GWAS as study sizes increase, that is as $v^*$ decreases, from within this range of selection coefficients. To this end, we first need to know the density of newly arising mutations at this range, which is

$$\rho(s) = n\, g(s)/n_c(v_0^*, s), \tag{3.12}$$

with $n$ the number of GWAS hits and $v_0^*$ the threshold contribution to variance in the current study. The expected number of variants discovered as GWAS increase is then

$$n(v^*) = \int_{s_{down}}^{s_{up}} \rho(s)\, n_c(v^*, s) \tag{3.13}$$

with $v^*$ the new threshold variance.

We can also predict the amount of variance explained by GWAS associations as study sizes increase from selection coefficients between $s_{down}$ and $s_{up}$. We define $v_c(v^*, s)$ as the amount of variance explained by a newly arising mutation with selection coefficients $s$ in a GWAS with threshold variance $v^*$. Therefore,

$v_c(v^*, s) = \int_{v>v^*, x>x^*} 2\mathrm{a}^2\mathrm{x}(1-\mathrm{x}) \cdot \mathrm{P}_c(x, a|s)$

and the expected amount of variance explained by GWAS is

$V(v^*) = \int_{s_{down}}^{s_{up}} \rho(s)\, v_c(v^*, s).$

Though these are only expectations, the variance around these expectations is negligible because of the large number of variants involved.

We estimate the error in $c$ and $\hat{\mathrm{g}}(s)$ and all other parameters calculated using them, such as the number of association in future GWAS and the variance they explain, by a nonparametric bootstrap. Specifically, we draw $n$ samples with replacement from the dataset and rerun the inference procedure. We bootstrap 100 times to obtain a collection of $c$ and $\hat{\mathrm{g}}(s)$ values. We then calculate confidence intervals for any value calculated from $c$ and $\hat{\mathrm{g}}(s)$. Repeating this process with many simulated datasets, we could verify that this bootstrap procedure produces well-calibrated confidence intervals for the mean and standard deviation of the logged selection coefficients (Figures A3.6A-B & A3.7A-B). Within the observed range of selection coefficients, we also obtain well-calibrated CI for the CDF of g(s) (Figures A3.6E-F & A3.7E-F). However, the CI for the probability density are only rough estimates of the true CI, even within the observable range (Figures A3.6C-D & A3.7C-D).

# Impact and Future Directions

The work presented in this thesis is very much a stepping stone for further theoretical and experimental analysis. It provides a thorough theoretical basis for the effects of demographic changes on mutation load and for the effects of pleiotropic stabilizing selection on the genetic basis of quantitative traits. This basis can be expended upon to address further factors and scenarios affecting mutation load and quantitative traits. In addition, the applications to data presented here can be applied to other human populations, other species and other traits. In the few years between the publication of Chapters 1 and 2 and the writing of this thesis, many fascinating extensions and applications of this work have already been published.

In Chapter 1 and Appendix 1, I present a detailed analysis of the effects of short term demographic events on mutation load. This analysis focused on the demographic changes seen in African and European human populations but can be used as the basis to analyze more dramatic demographic changes seen in other human populations and populations of other species. Indeed, some groups have already started using this analysis as the basis for addressing changes to load in archaic hominins and non-human species[1,2].

The analysis in Chapter 1 can also be used to address other questions. It has already been used to address the effects of differences in load on introgression between recently split populations[3] and to detect recessive effects on deleterious mutations[4]. It can be used as the basis to address further questions like the effects of recent environmental changes, like the advent of agriculture and modern medicine, on mutation accumulation in the human genome.

In Chapter 2 and Appendix 2, I presented a population genetic framework to interpret GWAS results for quantitative traits under pleiotropic stabilizing selection. My analysis includes many extensions to the basic pleiotropic stabilizing selection model. These extensions are taken in their small perturbation limit and further work is needed to address the effect of these extensions on the genetic basis of traits beyond this small perturbation limit. Of particular interest is the effect of adaptation on the genetic basis of traits[5], including large polygenic adaptation[6] and repeated local adaptations[7]. Another necessary avenue of investigation is extending this work to complex diseases[8], such as diabetes, multiple sclerosis and schizophrenia.

Only a sample of this work's applicability to data was presented in Chapter 2. In Chapter 2, we tested our results against GWAS for two traits but we intend to test it systematically against GWAS results for all available traits. In Chapter 3, we presented an extension of this work that should be able to directly address the effects of demography on the genetic basis of quantitative traits and estimate the selective effects on variants associated with traits.

Recent work has started addressing the biological mechanisms underlying our results. Our results are formulated from a purely theoretical, population genetic perspective without addressing the exact mechanisms leading to the observed pleiotropic stabilizing selection. Recent work has begun laying the theoretical framework for understanding the regulatory basis of quantitative trait variation[9,10] and detecting regulatory stabilizing selection effects[11]. I hope, in my future efforts, to contribute to these efforts.

# References

## Introduction


1. Rutherford, A. *A Brief History of Everyone Who Ever Lived: The Human Story Retold Through Our Genes*, (The Experiment, 2017).

2. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321-324 (2005).

3. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).

4. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010).

5. Henn, B.M., Cavalli-Sforza, L.L. & Feldman, M.W. The great human expansion. *Proceedings of the National Academy of Sciences*, 201212380 (2012).

6. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* **1**, 131 (2010).

7. DeGiorgio, M., Jakobsson, M. & Rosenberg, N.A. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proceedings of the National Academy of Sciences* **106**, 16057-16062 (2009).

8. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740-3 (2012).

9. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).

10. Casals, F. & Bertranpetit, J. Human Genetic Variation, Shared and Private. *Science* **337**, 39-40 (2012).

11. Lohmueller, K.E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-997 (2008).

12. Simons, Y.B., Turchin, M.C., Pritchard, J.K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat Genet* **46**, 220-224 (2014).

13. Simons, Y.B. & Sella, G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Current Opinion in Genetics & Development* **41**, 150-158 (2016).

14. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* **47**, 126-131 (2015).

15. Falconer, D.S. & Mackay, T.F.C. *Introduction to Quantitative Genetics*, 480 (Benjamin Cummings, Essex, England, 1996).

16. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5-22 (2017).

17. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics* **46**, 1173–1186 (2014).

18. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).

19. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).

20. Balding, D.J. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**, 781 (2006).

21. Ikram, M.K. *et al.* Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo. *PLOS Genetics* **6**, e1001184 (2010).

22. Manolio, T.A. A decade of shared genomic associations. *Nature* **546**, 360 (2017).

23. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45**, D896-D901 (2017).

24. Barton, N.H. & Turelli, M. Evolutionary quantitative genetics: how little do we know? *Annual review of genetics* **23**, 337–370 (1989).

25. Simons, Y.B., Bullaughey, K., Hudson, R.R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLOS Biology* **16**, e2002985 (2018).

26. Pickrell, J.K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* **48**, 709-17 (2016).

27. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).

## Chapter 1


1. Coventry, A. et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nature Communications **1**, 131 (2010).

2. Marth, G. T. et al. The functional spectrum of low-frequency coding variation. Genome Biology **12**, R84 (2011).

3. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science **336**, 740–743 (2012).

4. Nelson, M. R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science **337**, 100–104 (2012).

5. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science **337**, 64–69 (2012).

6. Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature (2012).

7. Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nature Genetics **39**, 1251–1255 (2007).

8. Wall, J. & Przeworski, M. When did the human population size start increasing? Genetics **155**, 1865–1874 (2000).

9. Voight, B. F. et al. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proceedings of the National Academy of Sciences of the United States of America **102**, 18508–18513 (2005).

10. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics **5**, e1000695 (2009).

11. Lohmueller, K. E. et al. Proportionally more deleterious genetic variation in European than in African populations. Nature **451**, 994–997 (2008).

12. Casals, F. & Bertranpetit, J. Human genetic variation, shared and private. Science **337**, 39–40 (2012).

13. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? The American Journal of Human Genetics **69**, 124–137 (2001).

14. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proceedings of the National Academy of Sciences **107**, 1752–1756 (2010).

15. Gibson, G. Rare and common variants: twenty arguments. Nature Reviews Genetics **13**, 135–145 (2012).

16. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature **461**, 747–753 (2009).

17. Schaffner, S. F. et al. Calibrating a coalescent simulation of human genome sequence variation. Genome Research **15**, 1576–1583 (2005).

18. Hartl, D. L. A Primer of Population Genetics (Sinauer Associates, Inc., 2000).

19. Travis, M. et al. Deleterious mutations can surf to high densities on the wave front of an expanding population. Molecular Biology and Evolution **24**, 2334–2343 (2007).

20. Lynch, M., Conery, J. & Burger, R. Mutational meltdowns in sexual populations. Evolution 1067–1080 (1995).

21. The 1000 Genomes Project Consortium. A map of human genome variation from population- scale sequencing. Nature **467**, 1061–1073 (2010).

22. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. Nature Methods **7**, 248–249 (2010).

23. Thornton, K. R., Foran, A. J. & Long, A. D. Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS Genetics **9**, e1003258 (2013).

24. Johnson, T. & Barton, N. Theoretical models of selection and mutation on quantitative traits. Philosophical Transactions of the Royal Society B: Biological Sciences **360**, 1411–1425 (2005).

25. Charlesworth, B. & Charlesworth, D. Elements of evolutionary genetics. (2010).

26. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. PLoS genetics **5**, e1000471 (2009).

27. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research **38**, e164–e164 (2010).

28. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols **4**, 1073–1081 (2009).

29. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. Genome research **19**, 1553–1561 (2009).

30. Schwarz, J. M., Rö delsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. Nature Methods 7, 575–576 (2010).

## Chapter 2


1. Falconer, D.S. & Mackay, T.F.C. *Introduction to Quantitative Genetics*, 480 (Benjamin Cummings, Essex, England, 1996).

2. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*, (Sinauer Sunderland, MA, 1998).

3. Provine, W.B. *The origins of theoretical population genetics: with a new afterword*, (University of Chicago Press, 2001).

4. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five Years of GWAS Discovery. *American Journal of Human Genetics* **90**, 7-24 (2012).

5. Barton, N.H. & Keightley, P.D. Understanding quantitative genetic variation. *Nature Reviews: Genetics* **3**, 11-21 (2002).

6. Manolio , T.A. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine* **363**, 166-176 (2010).

7. Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.-L.V. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* **494**, 234-237 (2013).

8. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

9. Zaitlen, N. *et al.* Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genet* **9**, e1003520 (2013).

10. Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-145 (2012).

11. Eichler, E.E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-450 (2010).

12. Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109**, 1193-1198 (2012).

13. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *American Journal of Human Genetics* **88**, 294-305 (2011).

14. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences* **111**, E455-E464 (2014).

15. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).

16. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics* **46**, 1173–1186 (2014).

17. Agarwala, V., Flannick, J., Sunyaev, S., Consortium, G.D. & Altshuler, D. Evaluating empirical bounds on complex disease genetic architecture. *Nature Genetics* **45**, 1418-1427 (2013).

18. Caballero, A., Tenesa, A. & Keightley, P.D. The Nature of Genetic Variation for Complex Traits Revealed by GWAS and Regional Heritability Mapping Analyses. *Genetics* **201**, 1601-1613 (2015).

19. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics* **69**, 124–137 (2001).

20. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences* **107**, 1752–1756 (2010).

21. de Vladar, H.P. & Barton, N. Stability and response of polygenic traits to stabilizing selection and mutation. *Genetics* **197**, 749-767 (2014).

22. Turelli, M. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theoretical Population Biology* **25**, 138-193 (1984).

23. Barton, N.H. & Turelli, M. Evolutionary quantitative genetics: how little do we know? *Annual review of genetics* **23**, 337–370 (1989).

24. Hill, W.G. & Kirkpatrick, M. What Animal Breeding Has Taught Us about Evolution. *Annual Review of Ecology, Evolution, and Systematics* **41**, 1-19 (2010).

25. Johnson, T. & Barton, N. Theoretical models of selection and mutation on quantitative traits. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1411-1425 (2005).

26. Lande, R. Natural selection and random genetic drift in phenotypic evolution. *Evolution* **30**, 314-334 (1976).

27. Hodgins-Davis, A., Rice, D.P. & Townsend, J.P. Gene Expression Evolves under a House-of-Cards Model of Stabilizing Selection. *Molecular Biology and Evolution* **32**, 2130-2140 (2015).

28. Byars, S.G., Ewbank, D., Govindaraju, D.R. & Stearns, S.C. Natural selection in a contemporary human population. *Proceedings of the National Academy of Sciences* **107**, 1787-1792 (2010).

29. Stulp, G., Pollet, T.V., Verhulst, S. & Buunk, A.P. A curvilinear effect of height on reproductive success in human males. *Behavioral Ecology and Sociobiology* **66**, 375-384 (2012).

30. Frederick, D.A. & Haselton, M.G. Why is muscularity sexy? Tests of the fitness indicator hypothesis. *Pers Soc Psychol Bull* **33**, 1167-83 (2007).

31. Endler, J.A. *Natural selection in the wild*, (Princeton University Press, 1986).

32. Kingsolver, J.G. *et al.* The strength of phenotypic selection in natural populations. *The American Naturalist* **157**, 245–261 (2001).

33. Charlesworth, B., Lande, R. & Slatkin, M. A Neo-Darwinian Commentary on Macroevolution. *Evolution* **36**, 474-498 (1982).

34. Barton, N.H. Pleiotropic models of quantitative variation. *Genetics* **124**, 773-782 (1990).

35. Kondrashov, A.S. & Turelli, M. Deleterious mutations, apparent stabilizing selection and the maintenance of quantitative variation. *Genetics* **132**, 603-18 (1992).

36. Wagner, G.P. Apparent Stabilizing Selection and the Maintenance of Neutral Genetic Variation. *Genetics* **143**, 617-619 (1996).

37. Pickrell, J.K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* **48**, 709-17 (2016).

38. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* **89**, 607-18 (2011).

39. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483-495 (2013).

40. Cotsapas, C. *et al.* Pervasive Sharing of Genetic Effects in Autoimmune Disease. *PLoS Genet* **7**, e1002254 (2011).

41. Andreassen, O.A. *et al.* Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate. *PLoS Genet* **9**, e1003455 (2013).

42. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-1241 (2015).

43. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540-2 (2012).

44. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-994 (2013).

45. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).

46. Visscher, P.M. & Yang, J. A plethora of pleiotropy across complex traits. *Nat Genet* **48**, 707-8 (2016).

47. Charlesworth, B. Evidence against Fisher's theory of dominance. *Nature* **278**, 848-849 (1979).

48. Segrè, D., DeLuna, A., Church, G.M. & Kishony, R. Modular epistasis in yeast metabolism. *Nature Genetics* **37**, 77 (2004).

49. Phadnis, N. & Fry, J.D. Widespread Correlations Between Dominance and Homozygous Effects of Mutations: Implications for Theories of Dominance. *Genetics* **171**, 385-392 (2005).

50. Wright, S. Fisher's theory of dominance. *Am Nat* **63**, 274-279 (1929).

51. Agrawal, A.F. & Whitlock, M.C. Inferences about the distribution of dominance drawn from yeast gene knockout data. *Genetics* **187**, 553-66 (2011).

52. Hill, W.G., Goddard, M.E. & Visscher, P.M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genet* **4**, e1000008 (2008).

53. Clayton, D.G. Prediction and Interaction in Complex Disease Genetics: Experience in Type 1 Diabetes. *PLoS Genet* **5**, e1000540 (2009).

54. Crow, J.F. On epistasis: why it is unimportant in polygenic directional selection. *Philos Trans R Soc Lond B Biol Sci* **365**, 1241-1244 (2010).

55. Ávila, V. *et al.* The Aaction of Stabilizing Selection, Mutation, and Drift on Epistatic Quantitative Traits. *Evolution* **68**, 1974-1987 (2014).

56. Wei, W.-H., Hemani, G. & Haley, C.S. Detecting epistasis in human complex traits. *Nat Rev Genet* **15**, 722-733 (2014).

57. Perry, J.R.B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92-97 (2014).

58. Scott, R.A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics* **44**, 991-1005 (2012).

59. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am J Hum Genet* **99**, 139-53 (2016).

60. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-569 (2010).

61. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**, 1114-1120 (2015).

62. Wood, A.R. *et al.* Another explanation for apparent epistasis. *Nature* **514**, E3-E5 (2014).

63. Evans, D.M. *et al.* Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet* **43**, 761-7 (2011).

64. The International Multiple Sclerosis Genetics Consortium. Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat Genet* **47**, 1107-1113 (2015).

65. Moutsianas, L. *et al.* Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nature genetics* **47**, 1107-1113 (2015).

66. Barton, N. & Turelli, M. Adaptive landscapes, genetic distance and the evolution of quantitative characters. *Genetical research* **49**, 157-173 (1987).

67. Bulmer, M.G. The genetic variability of polygenic characters under optimizing selection, mutation and drift. *Genetical research* **19**, 17-25 (1972).

68. Keightley, P.D. & Hill, W.G. Quantitative genetic variability maintained by mutation-stabilizing selection balance in finite populations. *Genetical research* **52**, 33-43 (1988).

69. Robertson, A. The effect of selection against extreme deviants based on deviation or on homozygosis. *Journal of Genetics* **54**, 236–248 (1956).

70. Wright, S. The analysis of variance and the correlations between relatives with respect to deviations from an optimum. *Journal of Genetics* **30**, 243-256 (1935).

71. Keightley, P.D. & Hill, W.G. Variation Maintained in Quantitative Traits with Mutation-Selection Balance: Pleiotropic Side-Effects on Fitness Traits. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **242**, 95-100 (1990).

72. Fisher, R.A. *The genetical theory of natural selection*, 272 (Clarendon Press, Oxford, England, 1930).

73. Lande, R. The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genet Res* **26**, 221-35 (1975).

74. Lande, R. The Genetic Covariance Between Characters Maintained by Pleiotropic Mutations. *Genetics* **94**, 203-215 (1980).

75. Ewens, W.J. *Mathematical Population Genetics 1: I. Theoretical Introduction*, (Springer Science & Business Media, 2004).

76. Martin, G. & Lenormand, T. A General Multivariate Extension of Fisher's Geometrical Model and the Distribution of Mutation Fitness Effects across Species. *Evolution* **60**, 893-907 (2006).

77. Sham, P.C. & Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* **15**, 335-46 (2014).

78. Berg, J.J. & Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLOS Genetics* **10**, e1004412 (2014).

79. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760-764 (2016).

80. Turchin, M.C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature genetics* **44**, 1015-1019 (2012).

81. Berg, J.J., Zhang, X. & Coop, G. Polygenic Adaptation has Impacted Multiple Anthropometric Traits. *bioRxiv* (2017).

82. Robinson, M.R. *et al.* Population genetic differentiation of height and body mass index across Europe. *Nat Genet* **47**, 1357-1362 (2015).

83. Jain, K. & Stephan, W. Rapid Adaptation of a Polygenic Trait After a Sudden Environmental Shift. *Genetics* **207**(2017).

84. Simons, Y.B., Turchin, M.C., Pritchard, J.K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat Genet* **46**, 220-224 (2014).

85. Lohmueller, K.E. The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. *PLOS Genetics* **10**, e1004379 (2014).

86. Wall, J.D. & Przeworski, M. When did the human population size start increasing? *Genetics* **155**, 1865-74 (2000).

87. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* **1**, 131 (2010).

88. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).

89. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925 (2014).

## Chapter 3


1. Falconer, D.S. & Mackay, T.F.C. *Introduction to Quantitative Genetics*, 480 (Benjamin Cummings, Essex, England, 1996).

2. Barton, N.H. & Turelli, M. Evolutionary quantitative genetics: how little do we know? *Annual review of genetics* **23**, 337–370 (1989).

3. Barton, N.H. & Keightley, P.D. Understanding quantitative genetic variation. *Nature Reviews: Genetics* **3**, 11-21 (2002).

4. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5-22 (2017).

5. Agarwala, V., Flannick, J., Sunyaev, S., Consortium, G.D. & Altshuler, D. Evaluating empirical bounds on complex disease genetic architecture. *Nature Genetics* **45**, 1418-1427 (2013).

6. Ewens, W.J. *Mathematical Population Genetics 1: I. Theoretical Introduction*, (Springer Science & Business Media, 2004).

7. Simons, Y.B., Bullaughey, K., Hudson, R.R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLOS Biology* **16**, e2002985 (2018).

8. Sawyer, S.A. & Hartl, D.L. Population Genetics of Polymorphism and Divergence. *Genetics* **132**, 1161-1176 (1992).

9. Boyko, A.R. *et al.* Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLOS Genetics* **4**, e1000083 (2008).

10. Halligan, D.L. *et al.* Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. *PLOS Genetics* **9**, e1003995 (2013).

11. Kim, B.Y., Huber, C.D. & Lohmueller, K.E. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics* (2017).

12. Ma, X. *et al.* Population Genomic Analysis Reveals a Rich Speciation and Demographic History of Orang-utans (Pongo pygmaeus and Pongo abelii). *PLOS ONE* **8**, e77175 (2013).

13. McManus, K.F. *et al.* Inference of gorilla demographic and selective history from whole-genome sequence data. *Mol Biol Evol* **32**, 600-12 (2015).

14. Sham, P.C. & Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* **15**, 335-46 (2014).

15. Kooperberg, C. & Stone, C.J. A study of logspline density estimation. *Computational Statistics & Data Analysis* **12**, 327-347 (1991).

16. Nelder, J.A. & Mead, R. A simplex method for function minimization. *The computer journal* **7**, 308–313 (1965).

17. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925 (2014).

18. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).

19. Bataillon, T. & Bailey Susan, F. Effects of new mutations on fitness: insights from models and data. *Annals of the New York Academy of Sciences* **1320**, 76-92 (2014).

20. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).

21. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* (2014).

22. Spain, S.L. & Barrett, J.C. Strategies for fine-mapping complex traits. *Human Molecular Genetics* **24**, R111-R119 (2015).

23. Albers, P.K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *bioRxiv* (2018).

24. Hansen, M.H. & Kooperberg, C. Spline Adaptation in Extended Linear Models (with comments and a rejoinder by the authors). *Statist. Sci.* **17**, 2-51 (2002).

Dear reader, thank you for reading this thesis! Please contact me at polygenicpizza@gmail.com to receive a small reward for your diligence. Yuval B. Simons

# Impact and Future Directions


1. Simons, Y.B. & Sella, G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Current Opinion in Genetics & Development* **41**, 150-158 (2016).

2. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* **47**, 126-131 (2015).

3. Harris, K. & Nielsen, R. The Genetic Cost of Neanderthal Introgression. *Genetics* **203**, 881-891 (2016).

4. Balick, D.J., Do, R., Cassa, C.A., Reich, D. & Sunyaev, S.R. Dominance of deleterious alleles controls the response to a population bottleneck. *PLoS Genet* **11**, e1005436 (2015).

5. Pritchard, J.K., Pickrell, J.K. & Coop, G. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Current Biology* **20**, R208-R215 (2010).

6. Stetter, M.G., Thornton, K. & Ross-Ibarra, J. Genetic architecture and selective sweeps after polygenic adaptation to distant trait optima. *bioRxiv* (2018).

7. Crawford, N.G. *et al.* Loci associated with skin pigmentation identified in African populations. *Science* **358**(2017).

8. Agarwala, V., Flannick, J., Sunyaev, S., Consortium, G.D. & Altshuler, D. Evaluating empirical bounds on complex disease genetic architecture. *Nature Genetics* **45**, 1418-1427 (2013).

9. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).

10. Liu, X., Li, Y.I. & Pritchard, J.K. Trans effects on gene expression can drive omnigenic inheritance. *bioRxiv* (2018).

11. Glassberg, E.C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J.K. Measurement of selective constraint on human gene expression. *bioRxiv* (2018).

# Appendix 1

## Contents

# 1 Model and simulations

Our basic model considers selection at a single site. We use the standard bi-allelic diploid model with (in this order) two-way mutation, viability selection, drift and, in some cases, migration [1]. Specifically, we assume there are two alleles at a site: normal (*N*) and deleterious (*D*). An *N* allele mutates to the *D* allele with probability $u$ per gamete, per generation and the reverse mutation occurs with probability $v$. Unless noted otherwise, we assume that mutation is symmetric, i.e., $u = v$. The absolute fitnesses of the three genotypes *NN*, *ND* and *DD* are $1$, $1 - hs$ and $1 - s$, respectively, where $s > 0$ and $h \geq 0$. We focus on semi-dominant ($h = \frac{1}{2}$) and fully recessive ($h = 0$) selection because these two cases exhibit the full range of qualitative behaviors (with selection acting primarily on heterozygotes in one and only on homozygotes in the other), but we also consider the robustness of our findings to other dominance coefficients (section 2.4). Allele frequencies in the next generation follow from Wright-Fisher sampling with these viabilities, sometimes with migration, and the population size and migration rates vary according to the demographic scenario considered.

For each demographic scenario, we ran simulations of a single site for the semi-dominant and recessive cases and varied the selection coefficient such that selection ranges from effectively neutral to strong. For a given set of parameters, the number of runs was determined by requiring a sampling error of less than 2% in estimates of the main summaries (e.g., the mean deleterious allele frequency and squared frequency). Error bars denoting estimates of one standard deviation around the mean are provided in all the graphs based on simulations, unless they are too small to be visible. Each run begins with one of the two alleles fixed, where the proportion of runs that start with each allele is given by the expectation at equilibrium. A burn-in period of $\geq 10N$ generations with constant population size $N$ follows in order to ensure an equilibrium distribution of segregating sites. The initial state is defined as ancestral and the other state as derived; the derived and deleterious allele frequencies are recorded at the end of the simulation. The code is written in C++ and is available upon request.

**Demographic scenarios.** We consider three demographic scenarios. The most detailed is the Out-of-Africa demographic model for African-Americans (AA) and European-Americans (EA) estimated by Tennessen et al. [2] (Supplementary Figure A1.1A). The model includes the Out-of-Africa split of European ancestors, changes in population size before and after the split (specifically a severe bottleneck in Europeans following the split and recent rapid growth in both Europeans and Africans) and migration between the populations after the split (see Supplementary Figure A1.1A for details). Finally, the model includes recent admixture between the populations, which

we include in our simulations only when we compare our results to data from AAs.

While the Tennessen et al. model was parameterized in a diffusion framework, i.e., in continuous time, Wright-Fisher simulations require discrete numbers of generations and individuals. We therefore divide the times by $25$ years per generation (the generation time that Tennessen et al. assume) and round the number of individuals associated with any of the parameters (e.g., growth) to the nearest integer. We implement migration by sampling alleles from the local population with probability $1 - m$ and from the other population with probability $m$ each generation.

We also study two simpler demographic scenarios. To understand the effects of recent 'explosive' growth of human populations, we use a simple model of exponential growth with parameters matching those of the African population in the Tennessen et al. model (see Supplementary Figure A1.1B for details). For the purpose of analysis, this scenario is sometimes extended by adding a period with constant population size after growth ends. Similarly, to investigate the effects of the bottleneck in Europeans at the Out-of-Africa split, we consider a simple model of a bottleneck with parameters matching those of the European bottleneck in the Tennessen et al. model (see Supplementary Figure A1.1C for details). Here, we sometimes extend the period after the reduction in population size to study longer-term equilibration to reduced population sizes.

**Validating the simulation.** We used two approaches to check the validity of the simulations. For a constant population size, we compared the frequency spectra from simulations with those expected under the diffusion approximation (cf. [3]) for the neutral case as well as for several semi-dominant and recessive selection coefficients (Supplementary Figure A1.6). We note that obtaining similar frequency spectra implies that simpler summaries, such as the number of segregating sites under neutrality or the average deleterious allele frequency at mutation-selection balance, will also be similar.

For the more elaborate Out-of-Africa demographic model, we compared the minor allele frequency spectrum from neutral simulations with the spectrum observed at non-coding sites in Fu et al. [4] We cosider non-coding sites for this purpose as these are assumed to be under the least selection (Supplementary Figure A1.7). In their Supplementary Figure 2A, Tennessen et al. find a close agreement between the observed spectra and a diffusion approximation under their demographic model. We find close agreement of our neutral simulations to data from both AAs and EAs and the slight differences that we do find are similar to those in their Supplementary Figure 2A [2].

**Sensitivity to mutation rate.** Unless noted otherwise, we follow Tennessen et al. [2] in using a mutation rate of $u = 2.36 \cdot 10^{-8}$ per bp per generation. Given that recent estimates suggest a

lower mutation rate (e.g. Kong et al. [5], Sun et al. [6]), we examine here the sensitivity of our simulation results to this assumption. We find the derived allele frequency spectrum to be extremely robust, remaining essentially unchanged when we double or halve the mutation rate (Supplementary Figure A1.8A). As expected, the number of segregating sites and the number of sites fixed for the derived allele increase (linearly) with the mutation rate (Supplementary Figure A1.8B). The increase in the number of sites fixed for the derived allele follows from the increased rate of fixation in the burn in period (akin to fixations that occur between the ancestor of humans and chimpanzees and the Out-of-Africa split). Thus, assuming a different mutation rate will affect some of our quantitative results. Notably, if the mutation rate in humans is indeed lower than the one we use, as recent estimates suggest, the proportion of segregating sites would be lower, resulting in an even smaller effect of recent demographic history on load than our analysis suggests (see section 2). Our qualitative finding of a negligible effect on load is unchanged. Moreover, our results concerning the effects of recent demography on genetic architecture derive from the frequency spectrum and therefore are unaffected.

## 2 The effects of demography on load

We assume that fitness is multiplicative across sites and that selected sites are at Linkage Equilibrium (LE). The absolute fitness of individual $i$ can then be written as

$$W_i = \prod_{j=1}^{M} w_{i,j},$$

where the product is taken over the $M$ sites contributing to fitness and $w_{i,j}$ is the contribution of site $j$, which depends on the genotype of the individual and on the selection and dominance coefficients at that site. Given LE, the contributions of sites to the expected fitness in the population are independent and therefore

$$E(W_i) = \prod_{j=1}^{M} E(w_{i,j}) \approx \exp(-\sum_{j=1}^{M}(2h_j s_j p_j q_j + s_j q_j^2)),$$

where $p_j$ and $q_j$ are the frequencies of the normal and deleterious alleles at site $j$. We note that the approximation applies for strong selection because the frequency $q_j$ is small, as well as for weak selection because then the selection coefficient is small. Finally, taking an expectation over evolutionary realizations (which is equivalent to an expectation over many sites with the same

parameters in a single realization) yields

$$E(W) \approx \exp(-\sum_{j=1}^{M}(2h_j s_j E(p_j q_j) + s_j E(q_j^2))). \tag{A1.1}$$

The latter expression relates the population dynamics at a site with the overall reduction in fitness.

Genetic load is defined as the relative reduction in average fitness caused by deleterious alleles, calculated as

$$L = \frac{W_{max} - \bar{W}}{W_{max}},$$

where $W_{max}$ is the fitness of an individual without deleterious alleles and $\bar{W}$ is the average fitness [1]. Denoting the terms associated with a single site in Equation A1.1 by

$$l(h,s) \equiv 2hsE(pq) + sE(q^2) = s(2hE(q) + (1-2h)E(q^2)), \tag{A1.2}$$

the fitness function can be rewritten as

$$E(W) \approx \exp(-\sum_{j=1}^{M} l(h_j, s_j)).$$

This form emphasizes that the reduction in fitness caused by a single site generally depends on the first two moments of the deleterious allele frequency. Specifically, in the semi-dominant model, it depends only on the first moment

$$l(\frac{1}{2}, s) = sE(q),$$

and in the recessive model it depends only on the second

$$l(0, s) = sE(q^2).$$

Moreover, this form shows that $l(h,s)$ provides a natural additive measure for the expected reduction in fitness caused by a site.

Throughout the manuscript we therefore use $l(h,s)$ as our measure for the contribution of a site to load. For a model with a single site, it coincides with the definition of load, as $E(L) = l(h,s)$. For more than one site,

$$E(L) \approx 1 - \exp(-\sum_{j=1}^{M} l(h_j, s_j)).$$

Given that in our model, the load from all sites is a simple function of the sum of $l(h,s)$ across sites, for brevity, we refer to $l(h,s)$ as load.

With a constant population size, the load exhibits three standard dynamic regimes depending on the scaled selection coefficient (Supplementary Figure A1.9): (i) An effectively neutral regime, in which $\alpha = 2Ns \ll 1$ and the effects of selection are negligible compared to drift; (ii) a weak selection (or nearly neutral) regime, in which $\alpha = 2Ns \approx 1$ and the effects of selection and drift are comparable; (iii) a strong selection regime, in which $\alpha = 2Ns \gg 1$ and selection dominates over drift.

In what follows our analysis is divided according to these three regimes. When the population size changes, the boundaries between regimes are affected. Moreover, the rate at which the equilibrium for a new population size is attained depends on the summary of the data considered. We consider summaries for segregating sites, e.g., the proportion of segregating sites and the allele frequency at these sites, and summaries for fixed sites, e.g., the proportion of sites fixed for the deleterious allele (which we call "fixed state"). Specifically, we are interested in the effects of demography on the contribution of segregating and fixed sites to load, which we refer to as "fixed" and "segregating" load, and in their sum, which we refer to as "total" load. We consider the behavior of these statistics for the two simple demographic models, which together allow us to understand all qualitative behaviors exhibited under the more detailed Tennessen et al. model (A1.10). For these demographic models, we primarily consider two modes of inheritance (semi-dominant and recessive).

To simplify our theoretical analysis, we make several reasonable assumptions about the parameters of the model. For brevity, we focus on the case with symmetric mutation ($u = v$) and, because we are considering human populations, we assume that the population mutation rate per site is small, i.e., that $\beta = 2Nu \ll 1$. We also assume that the selection coefficient is small, i.e., $s \ll 1$. A summary of our analyses are presented in Supplementary Figure A1.10 and Table A1.1. A detailed description of the behavior in each regime follows.

### 2.1 The effectively neutral regime

When selection is negligible compared to drift, the behavior of deleterious alleles is well approximated by that of neutral alleles. As the properties of neutral alleles (e.g., the proportion of segregating sites and frequency spectrum) in models with constant and varying population sizes have been studied exhaustively (e.g., [9, 10, 11]), here we focus only on the implications concerning load.

First, we consider how load depends on the selection coefficient at equilibrium for a constant population size. If deleterious alleles behave like neutral ones, the first two moments of the deleterious

allele frequency distribution do not depend on the selection coefficient and therefore the load is proportional to the selection coefficient (see Eq. A1.2). This explains the linear relationship between selection coefficient and load shown in Supplementary Figure A1.9.

At equilibrium, load depends negligibly on the population size. Using the diffusion approximation for the stationary deleterious allele frequency distribution [3], the expansion of the load to first order in $\alpha$ and $\beta$ yields

$$l(h,s) = \frac{s}{2}(1 - \frac{1}{2}\alpha - 2(1-2h)\beta).$$

Thus, as long as $\beta \ll 1$ and $\alpha \ll 1$, the load is well approximated by $s/2$ regardless of the population size and dominance coefficient (hence the similarity in load for the semi-dominant and recessive cases in Supplementary Figure A1.9). Intuitively, this follows from the fact that the great majority of sites are fixed, and because selection is negligible, half of them are fixed for the deleterious allele ($\frac{u}{u+v}$ for asymmetric mutation).

The same reasoning implies that changes in population size will have a negligible effect on the total load in this regime (Supplementary Figure A1.11). While changes in population size affect the proportion of segregating sites and thus their contribution to load, so long as the population mutation rate remains negligibly small ($\beta \ll 1$), the segregating load will remain negligible compared to the fixed load. In the bottleneck model, the proportion of segregating sites decreases to a new equilibrium after the reduction in population size (Supplementary Figure A1.11A). This explains the decrease in segregating load, which is balanced by an increase in fixed load (Supplementary Figure A1.10). By the same token, in the growth model, the segregating load increases but is balanced by a decrease in fixed load, resulting in a negligible change to the total load (Supplementary Figure A1.10 and Supplementary Figure A1.11B). In this case, however, segregating sites are still far from their new equilibrium at present (see the next section).

### 2.2 The weak selection regime

In the weakly selected regime, selection and drift have comparable effects on the dynamics of deleterious alleles. As a result, at equilibrium, even moderate differences in population size can affect the balance between selection and drift. Changes in population size also shift the balance, and are followed by transient changes at fixed and segregating sites until a new equilibrium is attained. To understand these effects, we consider the behavior at equilibrium and the rate at which it is approached. For this purpose, it is helpful to use the low mutation rate (LMR) approximation in which mutant alleles at a segregating site have a single origin; in other words, we ignore mutations

that arise during the sojourn of a mutant allele from the time it arises on a background fixed for the other allele to the time it reaches fixation or loss in the population.

**The effect of population size on the proportion of sites fixed for the normal and deleterious alleles.** At equilibrium, the rate at which deleterious alleles arise and fix is equal to the rate at which normal alleles arise and fix. This balance can be written as

$$2Nup\pi(-2Ns, h, \frac{1}{2N}) = 2Nvq\pi(2Ns, 1-h, \frac{1}{2N}),$$

where $\pi$ denotes the fixation probability, which depends on the scaled selection and dominance coefficients and on the initial frequency [12] (because $s \ll 1$, we ignore second order terms in $s$). For $s \ll 1$ and any dominance coefficient, this yields

$$\frac{q}{p} = \frac{u}{v}\frac{\pi(-2Ns, h, \frac{1}{2N})}{\pi(2Ns, 1-h, \frac{1}{2N})} \approx \frac{u}{v}e^{-2Ns}.$$

Namely, at equilibrium, the proportion of fixed deleterious sites declines exponentially with the scaled selection coefficient $\alpha = 2Ns$ (Supplementary Figure A1.12A). Thus, for a given selection coefficient $s$, the population size has a dramatic effect on the proportion of sites fixed for the deleterious allele, declining from the neutral, mutation-driven, proportions for $s \ll \frac{1}{2N}$ to approximately 0 for $s \gg \frac{1}{2N}$.

Importantly, however, when the population size changes, the new equilibrium proportion may be attained very slowly. The fractions, $p(t)$ and $q(t)$, of sites fixed for the normal and deleterious alleles $t$ generations after a change in population size (assuming $p(t) + q(t) = 1$) are well approximated by the model

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} -2N_a u\pi(-2N_a s, h, \frac{1}{2N_a}) & 2N_a v\pi(2N_a s, 1-h, \frac{1}{2N_a}) \\ 2N_a u\pi(-2N_a s, h, \frac{1}{2N_a}) & -2N_a v\pi(2N_a s, 1-h, \frac{1}{2N_a}) \end{pmatrix}\begin{pmatrix} p \\ q \end{pmatrix},$$

where $N_a$ is the population size after the change, and fixation times (on the order of $4N_a$ generations) are neglected. An additional contribution from sites that were segregating before the change is considered below. In this approximation, the change in the fraction of sites fixed for the deleterious alleles is

$$q(t) = q_a^{eq}\left(1 - e^{-\frac{t}{\tau}}\right) + q_b^{eq}e^{-\frac{t}{\tau}},$$

where $q_b^{eq}$ and $q_a^{eq}$ are the equilibrium fractions corresponding to the population sizes before and after the change, and

$$\tau = \left[2N_a\left(u\pi(-2N_a s, h, \frac{1}{2N_a}) + v\pi(2N_a s, 1-h, \frac{1}{2N_a})\right)\right]^{-1}$$

is the timescale of the exponential approach to the new equilibrium. For the semi-dominant case and $s \ll 1$, this time scale is well approximated by

$$\tau \approx \left[ u\frac{\alpha}{e^{\alpha}-1} + v\frac{\alpha}{1-e^{-\alpha}} \right]^{-1},$$

demonstrating that it is mutation-limited. This is also true for other dominance coefficients. In other words, following an instantaneous change in population size, the proportion of sites fixed for the deleterious allele will change extremely slowly, at a rate that is inversely proportional to the mutation rate (Supplementary Figure A1.12B and C).

Because the equilibrium is reached slowly, recent demographic changes in humans should have had little effect on the proportion of sites fixed for the deleterious alleles and hence on the fixed load. The bottleneck at the Out-of-Africa split is estimated to have reduced the population size from $\sim 14,000$ to $1,800$ approximately $2000$ generations ago [2]. Once a new equilibrium is reached, there will be a substantial increase in the proportion of fixed deleterious alleles; for example, for a semi-dominant deleterious allele with selection coefficient of $s = 10^{-4}$, it would increase it from $0.05$ to $0.4$. Yet the change over $2000$ generations is minimal, increasing this proportion only by $3 \cdot 10^{-5}$. The estimated $200$ generations since the onset of rapid growth in humans is similarly much too short a time period for any measurable effect on the fixed load (which in this case would decrease over large time periods).

**The effects of population size on segregating sites.** First we consider how the equilibrium properties of segregating sites depend on population size in models with constant population size (Supplementary Figure A1.13). The deleterious allele frequency at segregating sites decreases with increasing population size, because the efficacy of selection is greater in larger populations (Supplementary Figure A1.13A). In turn, the proportion of segregating sites increases with population size due to the (linear) increase in the number of mutations that enter the population every generation (Supplementary Figure A1.13B). This is true not only for the population as a whole but also for subsamples from it of any size (Supplementary Figure A1.13C). Finally, the deleterious allele frequency and proportion of segregating sites decrease with increasing dominance coefficient, as stronger selection in heterozygotes results in stronger selection on deleterious mutations (regardless of their frequency) and thus in a shorter sojourn through the population. Thus, in larger populations or if the dominance coefficient is greater, we expect a greater proportion of segregating sites with deleterious alleles at lower frequency.

The total load decreases monotonically when the population size increases (as can be shown using the stationary distribution based on the diffusion approximation [3], for example). This is not

true of the segregating load, because the increase in the mutational input can have a greater effect than the increase in the efficacy of selection (Supplementary Figure A1.13D). Indeed, for selection coefficients closer to neutrality, the increase in mutational input (and the proportion of segregating sites) dominates, causing the segregating load to increase with population size (akin to the behavior in the effectively neutral regime). In contrast, for selection coefficients closer to the strong selection regime, the increase in the efficacy of selection dominates, leading to a reduction in segregating load (akin to the stronger selection regime; see section 2.3).

Next we consider the effects of a change in population size. We begin by noting that, for a given population size, the expected sojourn time of deleterious and beneficial mutations that reach fixation is shorter than that for a neutral mutation and is thus on the order of $4N$ generations or less [3]. This implies that on the order of $4N_a$ generations after a change in population size, most of the *old mutations* (i.e., those that segregated before the population size changed) have been absorbed (either due to loss or fixation), and replenished by *new mutations* (that arose and spread through the population at its new size). When this turnover process is complete, new segregating sites approach their equilibrium proportions (given a background of fixed sites).

In the bottleneck model, the reduction in the efficacy of selection causes an increase in total load, where the behavior of the components of load can be understood as follows (Supplementary Figure A1.2). Focusing first on the contribution of old mutations to the fixed load: When old mutations are absorbed, the reduction in the efficacy of selection leads more deleterious alleles to fix than would have had the population size remained constant (at the larger size), eventually resulting in an increase in fixed load. The increase can be approximated by

$$\Delta(s,h,u,N_b,N_a) = \int_0^1 \left(\pi(-2N_a s,h,x) - \pi(-2N_b s,h,x)\right) f(x;h,2N_b s,2N_b u)dx,$$

where $f(x;h,2N_b s,2N_b u)$ is the stationary distribution before the change in population size [3]. The increase is maximized for selection coefficients at which the change in population size leads selection to transit from strong to weak, and is negligible outside this range (Supplementary Figure A1.2A; explaining why it is more pronounced in Supplementary Figure A1.2C and D than in E and F, correspondingly). The increase in deleterious fixations and load is then followed by a long-term, slower increase in the fixed load due to new mutations (Supplementary Figure A1.2C-F). In the parameter regime where the fixation of old mutations makes a substantial contribution to load, there is also a transient increase in segregating load before the mutations fix (in Supplementary Figure A1.2C for example). These effects are more pronounced in the recessive case, because of the greater frequency and proportion of segregating sites. Now focusing on the segregating load (Supplementary Figure A1.2B): when segregating sites attain equilibrium, the reduction

in population size causes a decrease in segregating load for lower selection coefficients (Supplementary Figure A1.2C and D) and an increase for higher selection coefficients (Supplementary Figure A1.2E and F). Thus, for higher selection coefficients in the weak selection range, both old and new mutations contribute to the transient increase in segregating load observed in Supplementary Figure A1.10. For the lower selection coefficients in this range, the segregating load decreases both in the short and long term but the fixation of old mutations still results in an overall increase to the total load (Supplementary Figure A1.10). Importantly, however, on the timescale estimated for the bottleneck at the Out-of-Africa split (vertical line in Supplementary Figure A1.2), these effects amount to a tiny increase in total load (Supplementary Figure A1.10).

What about in the case of growth? Human population growth is thought to have started a couple hundred of generations ago, ending with an effective population size in the hundreds of thousands and starting from a size that was thirty-fold smaller [2]. Given the estimated growth parameters, there was insufficient time for the deleterious alleles that segregated before the onset of growth to change their frequencies substantially. Indeed even with the increase in the efficacy of selection as the population size increases, in this regime, selection is too weak to have caused a substantial change in allele frequency over hundreds of generations (although it could have caused the absorption of very rare or very high frequency alleles). After growth, the resulting frequency spectrum of deleterious alleles thus reflects a superposition of the spectrum of segregating sites before growth and of the spectrum at the large number of sites in which mutations were introduced after the onset of growth (Supplementary Figure A1.14). The many new mutations remain at low frequencies. Because of an increase in the proportion of segregating sites, the segregating load increases at the expense of fixed load, but with negligible effects on the total load, given both the low frequency of new mutations as well as the opposing contributions of normal and deleterious mutations (Supplementary Figure A1.10).

### 2.3 The strong selection regime

In this regime, purifying selection is sufficiently strong to prevent deleterious alleles from reaching high frequencies, let alone fixation. It follows that there is only segregating load. If we assume that the deleterious allele frequency is small and that the dominance coefficient is sufficiently large, then the load is well approximated by

$$l(h, s) \approx 2hsE(q).$$

Stated another way, when selection against heterozygotes is sufficiently strong, then deleterious homozygotes would be too rare to affect load. Under these assumptions, the diffusion approximation at equilibrium with a constant population size [3] yields

$$E(q) \approx \frac{u}{hs},$$

implying that the load is well approximated by

$$l(h, s) \approx 2u.$$

We refer to the cases where these conditions are met as quasi-dominant.

In the recessive case, the load depends on the second moment of deleterious allele frequency. Assuming once again that the deleterious allele frequency is small, the diffusion approximation at equilibrium with a constant population size [3] yields

$$E(q^2) \approx \frac{u}{s},$$

implying that the load is well approximated by

$$l(0, s) \approx u.$$

The expressions for load in both cases are identical to the classic ones for mutation-selection balance, which are derived assuming an infinite population size [12]. They imply that at equilibrium, the load depends neither on the selection coefficient (explaining the plateaus in Supplementary Figure A1.9) nor on the population size.

When the dominance coefficient is sufficiently small, however, the load does depend on population size (Supplementary Figure A1.15). This will be the case when selection against heterozygotes is weak, i.e. when $2Nhs \gg 1$ does not hold, as then both moments of deleterious allele frequency make comparable contributions to load. Holding the selection coefficient and population size constant, in this range of dominance coefficients, the load varies continuously with $h$ between $u$ and $2u$ (Supplementary Figure A1.15A). In turn, holding $h \ll 1$ and $N \gg 1$ constant, increasing $s$ also leads the load to vary from $u$ to $2u$ (Supplementary Figure A1.15B).

Next, we consider the effect of changes in population size, for the quasi-dominant and then the recessive case. We show that in the quasi-dominant case, the load remains constant and is well approximated by the classic derivations for mutation-selection balance. In the recessive case, the load exhibits transient changes before it returns to its equilibrium level.

**The quasi-dominant case**

In the quasi-dominant case, we can assume deleterious alleles are sufficiently rare that selection against deleterious homozygotes can be ignored and selection has negligible effects on average fitness. Under these conditions, we can approximate the trajectory of a deleterious allele using a branching process (cf. [13]), in which the number of copies that a given deleterious allele gives rise to in the next generation follows a distribution that is independent on the frequency of deleterious alleles in the population.

Consider a single deleterious allele that was introduced by mutation at time $t = 0$ and denote by $Z(t)$ the number of deleterious alleles that it gives rise to at generation $t$. The number of mutant alleles in the next generation can then be expressed as

$$Z(t+1) = \sum_{i=1}^{Z(t)} X_i(t),$$

where $X_i(t)$ denotes the number of offspring of the $i$'th allele at time $t$ and $i = 1, \ldots, Z(t)$. We denote the expected number of offspring of a single allele by $\lambda$, i.e., $E(X_i(t)) = \lambda$; if we ignore mutations back to the beneficial allele then $\lambda = 1 - hs$ and if we include them then $\lambda = 1 - hs - v$. The expected number of alleles in the next generation is then

$$E(Z(t+1)) = E(\sum_{i=1}^{Z(t)} X_i(t)) = \sum_{j=1}^{\infty} Pr(Z(t) = j) j E(X_i(t)) = E(Z(t))\lambda, \tag{A1.3}$$

or

$$E(Z(t)) = \lambda^{\tau}. \tag{A1.4}$$

Now consider the expected number of deleterious alleles at mutation-selection balance. For this purpose, we measure time backwards from the present. We denote by $Y_\tau(\tau)$ the number of mutations introduced $\tau$ generations ago and by $Y_\tau(t)$ the number of alleles that they give rise to at time $t$. The number of deleterious alleles at the present can then be expressed as the sum of contributions from all the mutations in the past, i.e. $\sum_{\tau=1}^{\infty} Y_\tau(0)$, where, from Equation A1.4,

$$E(Y_\tau(0)) = Y_\tau(\tau)\lambda^{\tau}.$$

In turn, the expected number of new mutations in a given generation is well approximated by

$$E(Y_\tau(\tau)) = 2Nu.$$

It follows that the expected deleterious allele frequency is

$$E(q) = \frac{1}{2N} E(\sum_{\tau=1}^{\infty} Y_\tau(0)) = \frac{1}{2N} \sum_{\tau=1}^{\infty} E(Y_\tau(\tau))\lambda^{\tau} = \frac{u}{hs},$$

and thus the expected contribution to load is $2u$ - well-known results for mutation-selection balance.

Next, we consider a changing population size. We denote by $N(t)$ the population size $t$ generations in the past and by $a(t) = \frac{N(t-1)}{N(t)}$ the proportional change in one generation. Now the expected number of new mutations introduced at a given time is proportional to the population size

$$E(Y_\tau(\tau)) = 2N(\tau)u,$$

but the fraction of new mutations in the population remains constant ($u$). Similarly, the expected number of alleles in the next generation is affected by changes in population size

$$E(Y_\tau(t-1)) = \lambda a(t) E(Y_\tau(t)),$$

but their fraction is not, because their increase in number is precisely offset by the increase in population size

$$E(\frac{Y_\tau(t-1)}{2N(t-1)}) = \lambda a(t) \frac{N(t)}{N(t-1)} E(\frac{Y_\tau(t)}{2N(t)}) = \lambda E(\frac{Y_\tau(t)}{2N(t)}).$$

It follows that the proportional contribution of alleles to the present is the same as that in a constant population size:

$$E(\frac{Y_\tau(0)}{2N(0)}) = u\lambda^\tau,$$

leaving the deleterious allele frequency and the load at the present unchanged (at $\frac{u}{hs}$ and $2u$). In other words, the expected frequency of deleterious alleles and therefore the load follow the same deterministic dynamic as they do in a population of constant size, because when the population size changes, the increase (decrease) in the copy number is precisely offset by the increase (decrease) in population size.

We note that incorporating reverse mutation and migration will not change this conclusion. Reverse mutation would reduce $\lambda$, while introducing migration would be similar to both decreasing $\lambda$ (due to migration of deleterious alleles out of the population) and increasing the mutational input (due to migration of deleterious mutations into the population).

Our results clarify how the expected deleterious allele frequency and proportion of segregating sites at equilibrium depend on population size. When the population mutation rate is sufficiently low, a site switches intermittently between having no deleterious alleles and having a single mutation (by origin) in the population (Supplementary Figure A1.16A). Under these conditions, in a larger population size, the mutational input is larger and thus the proportion of time that a site

is segregating increases (Supplementary Figure A1.16B). Because the trajectory of a mutation in terms of numbers of copies does not depend on the population size, the frequency of the mutation is proportional to $1/N$, so the expected frequency of deleterious alleles at segregating sites scales with $1/N$ (Supplementary Figure A1.16C). In turn, when the population mutation rate is sufficiently high, deleterious alleles are almost always present and often have several mutational origins. Under these conditions, the proportion of segregating sites approaches $1$ (Supplementary Figure A1.16B). Given that the expected frequency at segregating sites is $x = \frac{q}{S_{2N}}$, it follows that the allele frequency asymptotes to $q = \frac{u}{hs}$ (Supplementary Figure A1.16C). In turn, the variance in allele frequency decreases with population size and asymptotes to $0$ in the infinite population size limit.

After a change in population size, a new equilibrium is attained much more rapidly in the strong selection regime because of the rapid turnover of deleterious alleles (see Supplementary Figure A1.17). However, load is unaffected.

Thinking in terms of the branching process helps us to evaluate previous conjectures about the possible effects of human growth on deleterious alleles. For example, Keinan and Clark [15] suggest that "Some degree of genetic risk for complex disease may be due to this recent rapid expansion of rare variants in the human population". It is indeed the case that the expected copy number of deleterious alleles should be greater under exponential growth; specifically, for a population growing at a geometric rate $\gamma$ per generation, the copy number will change at a geometric rate of $\lambda + \gamma$ per generation, which will result in an increase if $\lambda + \gamma > 1$. Moreover, population growth increases the sojourn time of a deleterious mutation and, when $\lambda+\gamma > 1$, there is a finite probability it would never go extinct [16]. Importantly, however, the expected *frequency* of quasi-dominant deleterious alleles remains constant, so human population growth has no effect on load.

**The recessive case**

In this case, the load at equilibrium is again insensitive to population size, but the underlying reasons are quite different than in the quasi-dominant case. In the recessive model, a deleterious allele behaves neutrally while at low frequencies. As a result, its sojourn time (i.e., the expected time that it spends at frequency $x$) is well approximated by that of a neutral allele (Supplementary Figure A1.18B). When the frequency $x$ reaches $2Nsx^2 \approx 1$, selection on homozygotes for the deleterious alleles kicks in, and the allele should spend little time above this frequency. In the low mutation rate (LMR) approximation, we can therefore approximate the sojourn time of a recessive

deleterious allele as

$$\tau(x) \approx \begin{cases} \frac{2(2N-1)}{1-x} & \text{if } 0 \leq x \leq \frac{1}{2N} \\ \frac{2}{x} & \text{if } \frac{1}{2N} \leq x < \frac{1}{\sqrt{2Ns}} \\ 0 & \text{if } \frac{1}{\sqrt{2Ns}} \leq x < 1 \end{cases},$$

where the expressions for $x < 1/\sqrt{2Ns}$ are the sojourn times (in generations) for a neutral allele (Fig A1.18B). In this approximation, the expected contribution of a deleterious mutation to load is then

$$s\int_0^1 x^2\tau(x)\mathrm{d}x \approx s\int_0^{\frac{1}{\sqrt{2Ns}}} x^2\frac{2}{x}\mathrm{d}x = \frac{1}{2N},$$

and, given that the expected input of new mutations per generation is $2Nu$, the overall expected load is

$$l(0,s) \approx 2Nu\frac{1}{2N} = u.$$

In other words, (in the low mutation limit) for a given population size $N$, a recessive allele behaves neutrally up to a frequency of $N^{-\frac{1}{2}}$, resulting in an expected contribution to load that is proportional to $N^{-1}$. In turn, the mutational input is proportional to $N$, so they exactly offset.

This back of the envelope approximation also provides an intuitive explanation for the way in which the properties of segregating sites at equilibrium depend on population size (Fig A1.18). First, we consider the proportion of segregating sites (Fig A1.18A). When the population size is sufficiently small for the LMR approximation to apply, the proportion of segregating sites can be approximated by the ratio of the sojourn time of a single mutant through the population to the time between appearances of mutations, namely:

$$S_{2N} \approx \frac{\int_0^1 \tau(x)\mathrm{d}x}{\frac{1}{2Nu}} \approx 2Nu(ln(2N/s)+2).$$

In a larger population size and hence with a larger mutational input, mutations of different origin will overlap, resulting in a slower increase in the proportion of segregating sites with population size. When the mutational input becomes sufficiently large, this proportion asymptotes to 1. Next, we consider the frequency of deleterious alleles. In the LMR approximation, the frequency spectrum of segregating sites can be approximated using the neutral sojourn times up to the threshold frequency $\frac{1}{\sqrt{2Ns}}$ (Fig A1.18B), yielding an average frequency of $E(x) \approx \frac{\frac{2}{\sqrt{2Ns}}}{2+ln\frac{2N}{s}}$. As the population size increases, such that mutations of different origins overlap, the decrease in average frequency becomes slower and asymptotes to $E(x) = E(q) = \sqrt{u/s}$ (Fig A1.18C). Lastly, the turnover time of segregating sites for a given population size $N$ is on the order of $2\sqrt{\frac{2N}{s}}$. As it was for other regimes, this is the time scale for the process of equilibration following a change in population size.

We now consider the implications for the bottleneck and growth models. In the bottleneck model, after the reduction in population size, there is an increase in load followed by a decrease back to the equilibrium level (Supplementary Figure A1.19A). The transient increase in load (blue arrow in Supplementary Figure A1.19A) is dominated by the contribution of mutations that segregated before the decrease in population size. The proportion of sites that segregated before was greater and their frequencies lower than after the population size reduction, and while these segregating mutations are gradually absorbed, some of them will drift to higher frequencies, generating a transient surge in load (Supplementary Figure A1.19B). In turn, the newly introduced mutations have yet to reach equilibrium frequencies and, given that the contribution of the lower frequencies to load is much smaller, they contribute negligibly. In the Tennessen et al. model, the time that elapsed since the bottleneck is longer and the segregating sites are therefore closer to the new equilibrium (green arrow in Supplementary Figure A1.19A). Correspondingly, the relative contribution of new mutations is greater and their frequency distribution is closer to equilibrium with the new population size, and yet some contribution from the older mutations remains (Supplementary Figure A1.19C). These considerations also explain why load exceeds above equilibrium levels in the strong selection regime in Supplementary Figure A1.10.

In the growth scenario, we see the opposite transient effect: the load is reduced before recovering to its equilibrium level (Supplementary Figure A1.19D). After the growth period, the number of segregating sites is greatly increased, but the new mutations have had little time to drift to higher frequency. As a result, new mutations segregate at very low frequencies and contribute negligibly to load (Supplementary Figure A1.19E and F). In turn, mutations that segregated before growth have decreased in frequency due to the increased efficacy of purifying selection, and so their contribution to load declines substantially (Supplementary Figure A1.19E and F). The result is a transient reduction in load (seen in Supplementary Figure A1.10 as well as in Supplementary Figure A1.19D).

### 2.4 Models with dominance coefficients other than $0$ and $\frac{1}{2}$

Here we provide summaries of simulations with dominance coefficients other than $0$ and $1/2$ to illustrate that the same qualitative behaviors are observed. As shown in Supplementary Figure A1.20, all of the observed qualitative behaviors are included in our previous analysis and summarized in Table A1.1, with one possible exception.

The exception is in the bottleneck model in cases with dominance coefficients $h > 1/2$, where the total load is reduced for lower selection coefficients in the weak selection regime. The reason for

this reduction in load is analogous to that for the increase in load that we saw in the recessive case in the same selection regime. For dominance coefficients greater than half, the extinction of low frequency deleterious alleles that segregated before the reduction in population size decreases load more than the fixation of high frequency deleterious alleles increases it. The opposite is true for dominance coefficients smaller than half.

## 3 Data analysis and interpretation

We used data from Fu et al. (2012) [4] and from the 1000 Genomes Project [8]. Allele frequency estimates from Fu *et al.* are available from the NHLBI GO Exome Variant Server (`http://evs.gs.washington.ed`). These provide estimates of the derived allele frequencies at exonic SNVs in European- and African-Americans (EA and AA). Variants with allele frequencies 0 or 1 in both EA and AAs were excluded.

The haploid sample sizes in Fu et al were EA Autosomal: 8596, EA X: 6717, AA Autosomal: 4434, AA X: 3852. Our primary analysis in the main paper (reported in Figure 3) uses the full sample sizes with the autosomal data. For the purpose of Table A1.2 we wished to compare means on the X and autosomes. Since mean allele frequencies of segregating sites are affected by total sample size, we implemented the following subsampling strategy to facilitate direct comparisons between X and autosomes. First, we converted the reported allele frequencies for each site back into allele counts (i.e., multiplying each reported frequency by the relevant haploid sample size). Next, we randomly subsampled the autosomal EA and AA variants and the X chromosome EA variant allele frequencies down to a sample size of 3852 chromosomes each, in order to match the haploid sample size for the African-American X chromosome. Subsampling was done without replacement, using the hypergeometric sampling function in R. After sub-sampling, variants whose allele frequencies were both either 0 or 1 were once again dropped. Two-sided t-tests were used to test for allele frequency differences between groups.

1000 Genomes Project vcf files (Phase 1 Version 3) were downloaded from the official 1000 Genomes public server (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/`). YRI and CEU individuals with (at least) exome sequencing coverage were extracted from the original .vcf files (88 YRI individuals and 81 CEU individuals). 7 YRI individuals, chosen at random, were removed to match sample sizes between YRI and CEU. Variants that were fixed for either allele in both populations were removed. Any variant that was not an SNV or did not contain ancestral allele information was also dropped.

A natural measure for comparing the difference in load between two populations is to count the mean number of derived alleles per individual at SNVs segregating within the joint sample. Note that it is essential in these calculations to define SNVs using the joint sample, otherwise sites that are fixed for the derived allele in Population A but not in Population B would lead to the erroneous conclusion that there are more derived alleles in B than in A.

For our analysis, we found that it is convenient to work with the mean derived allele frequency within each functional class. This quantity allows us to compare frequencies directly between classes, and is also conveniently computed from the Fu et al frequency data. These two measures (mean derived frequency and number of derived alleles per individual) are proportional to one another and hence must yield identical conclusions about the relative load in different populations (for a given functional class: DAF multiplied by twice the number of SNVs yields the number of derived alleles per individual, assuming that missing data have been filled in appropriately). Notice also that we are dealing with mean numbers of alleles, and so these measures are unaffected by deviations from HWE or LE which affect the variance in numbers of derived alleles per individual but not the means.

Of course the number of derived alleles is not equivalent to the number of deleterious alleles, as some variants may be neutral; additionally for weakly selected sites there is a small probability at each site that the ancestral allele is deleterious. Nonetheless, the load is expected to be monotonically increasing with the number of derived alleles. As shown in Supplementary Figure A1.3, we predict that at semidominant sites there should be essentially no difference in mean derived frequency between AAs and EAs, regardless of selection coefficient. At recessive sites we would expect a small increase in mean frequency in AAs at moderately and strongly selected sites. The fact that we do not observe any significant difference in allele frequencies at "probably damaging" sites argues that the majority of these sites are at least partially dominant.

Mean derived allele frequencies were calculated for both populations at autosomal noncoding, synonymous, and nonsynonymous sites, as well as autosomal nonsynonymous variants belonging to the different functional categories. Standard errors for each category were estimated using the standard deviation in DAF across sites, divided by the square root of the number of sites in that category. For individual-level analyses, we computed the SD in mean number of variants per individual by bootstrapping across sites. The bootstrap analysis accounts for the evolutionary sampling variance in allele frequencies.

The ANNOVAR suite of scripts [21] was used to obtain functional predictions for each SNP from each of four prediction methods: PolyPhen2 [22], SIFT [23], LRT [25] and MutationTaster [24].

Default program settings were used in each case. The functional designations for each program are as follows: PolyPhen2: D (Probably Damaging), P (Possibly Damaging), B (Benign). SIFT: D (Damaging), T (Tolerant), LRT: D (Deleterious), N (Neutral) and U (Unknown). MutationTaster: A (Disease Causing Automatic), D (Disease Causing), P (Polymorphism Automatic) and N (Polymorphism). Coding versus non-coding and synonymous versus non-synonymous designations were also determined using ANNOVAR. (Note that we also tested the SeattleSeq annotations, and found that the overall numbers were similar (though not identical) to those obtained from ANNOVAR; as with ANNOVAR we found no evidence for a difference in DAF between populations.)

We observed that a strong reference bias exists at sites for which the genome reference sequence carries the derived reference allele. This bias has also been observed by David Reich and Shamil Sunyaev (personal communication). All four functional prediction programs designate a very high proportion of these sites as being likely nonfunctional or benign, even when the reference allele is rare in the population overall. When we condition on the overall population frequency at these sites, we find that a given site is much more likely to be classified as a probably damaging site if the reference genome carries the ancestral allele than if it carries the derived allele (Supplementary Figure A1.4).

To deal with this bias, we treated the functional designations at sites where the reference allele is derived as unreliable. As an alternative, we binned all SNVs into a series of allele frequency bins (i.e., the bins shown in Supplementary Figure A1.4). We assumed that when we condition on the population allele frequency in a very large sample (i.e., the Fu et al sample) that the identity of the genome reference allele carries essentially no further information about the likely functional properties of a variant. Thus, within a bin, the fraction of derived-reference SNVs that fall into each functional category can be predicted from the fraction of ancestral-reference SNVs in that functional category. Thus for example, if 20% of the ancestral-reference SNVs in a given bin have functional category X, then we assume that each of the derived-reference SNVs in that bin has a 20% probability of also being in functional category X. The mean frequency of all SNVs in category X is estimated by summing across all ancestral-reference SNVs in category X plus a sum of contributions from all derived-reference SNVs, weighted by the estimated probabilities that each is in X. As shown in Table A1.3, the bias correction makes a substantial difference to the data analysis. Prior to applying the bias correction, the mean frequency in AAs is substantially higher than in EAs (presumably because more than half of the reference genome sequence is of non-African origin (Supplement of [14], p145)), but the bias correction makes the two frequencies virtually identical as predicted for models with dominance.

We also provide supplementary results in which we made use of a new unpublished version of PolyPhen's PSIC scores that are calculated in a human-independent (i.e., unbiased) manner. (Thanks to Ivan Adzhubey and Shamil Sunyaev for access to these.) These produce results that are very similar to those from our bias-corrected version, in the sense of showing no difference between populations.

# 4 The effects of demography on the genetic architecture of disease risk

A great deal of interest focuses on understanding how recent demographic history has affected the genetic architecture of disease and specifically whether the recent explosive growth has increased the contribution of rare variants to disease risk [17, 15, 18, 2]. Here, we use the theory that we developed to elucidate some of these effects. *Note that while in what follows we refer to disease risk, it also applies to any other quantitative trait.*

## 4.1 A model relating allele frequencies to disease susceptibility

We first consider the relationship between selection on individual loci and disease risk. The few models for this relationship differ sharply in their assumptions. At one extreme, Pritchard [19] assumed that variants that increase disease susceptibility tend to be deleterious, but that otherwise there is no relationship between the strength of selection acting on these loci and the extent to which they increase disease susceptibility. In turn, Eyre-Walker [20] assumed a correlation between the strength of selection at a locus and its contribution to disease susceptibility. All else being equal, a stronger relationship between the disease risk and fitness implies that the variants that contribute more to disease risk are under stronger selection and, as a result, tend to be younger and rarer. It also follows that their frequency distribution would be more susceptible to the effects of recent demographic events. Here we consider models for the two extremes: one in which the effect sizes are independent on the selection coefficients and the other where the effect sizes are proportional to the selection coefficients.

To model how genetic variation relates to disease risk, we consider the $L$ loci that contribute to disease risk and denote the genotype of individual $i$ at these loci by $\mathbf{G_i} = (g_{i,1}, \ldots, g_{i,L})$. We assume that each of the loci is bi-allelic, with a normal ($N$) and susceptible ($S$) alleles, and therefore denote the genotype at locus $j$ ($j = 1, \ldots, L$) as $g_{i,j} = NN$, $NS$, or $SS$. We then assume that the

probability of developing the disease (ignoring life-history details) takes the form

$$P(\mathbf{G}) = F(\sum_{j=1}^{L} \alpha_j(g_j)),$$

where $F$ is a monotonically increasing function with continuous derivatives that takes values between 0 and 1 and that

$$\alpha_j(g) = \begin{cases} 0 & \text{if } g = NN \\ h_j a_j & \text{if } g = NS \\ a_j & \text{if } g = SS \end{cases},$$

where $h_j$ and $a_j$ denote the dominance coefficient and effect size of the contribution to susceptibility at locus $j$. Finally, we assume that the effect of each locus is small, such that we can approximate the variance in susceptibility by *the first term in a Taylor expansion, i.e.,*

$$V(P(G)) \approx [F'(\sum_{j=1}^{L} E(\alpha_j(g_j)))]^2 \sum_{j=1}^{L} V(\alpha_j(g_j)), \tag{A1.5}$$

where the variances are taken over the population and

$$V(\alpha(g); x, a, h) = a^2 x(1-x)\left[(2h-1)^2 x^2 + (1-4h^2)x + 2h^2\right],$$

*where $x$ is the $S$-allele frequency.*

*Our model in which the effect sizes are independent on the selection coefficients (and similarly for dominance coefficients) follows directly.* For simplicity we assume that the effect sizes and dominant coefficients are constant, as assuming a distribution yields similar results for all the quantities that we consider below. *The variance in disease susceptibility then follows from Eq. A1.5, where the $a_j$'s and $h_j$'s are constant across loci and the distribution of allele frequencies (the $x$'s) is determimined by the (independent) selection and dominance coefficients (for fitness) at these loci.*

*Next, we consider the model in which the disease itself is the agent of selection. In other words that the fitness cost results entirely from the probability of developing the disease.* Denoting the fitness of affected individuals by $W_a$ and of unaffected by $W_u$, the relationship between fitness, $W$, and the probability of developing the disease then takes the form

$$W = PW_a + (1-P)W_u.$$

In turn, in our model, the relationship between genotype and fitness is

$$W(\mathbf{G_i}) = \prod_{j=1}^{L} w_{i,j} \approx \exp\left(-\sum_{j=1}^{L} \alpha_j(g_{i,j})\right),$$

where

$$\alpha_j(g) = \begin{cases} 0 & \text{if } g = NN \\ h_j s_j & \text{if } g = ND \\ s_j & \text{if } g = DD \end{cases},$$

and we assume that $s_j \ll 1$ and therefore use an exponential approximation. Equating our two expressions for fitness leads to the following model for the relationship between disease risk and genotype

$$P(\mathbf{G}) = \frac{W_u - W(\mathbf{G})}{W_u - W_a} = \frac{W_u}{W_u - W_a} - \frac{1}{W_u - W_a} exp\left(-\sum_{j=1}^{L} \alpha_j(g_j)\right).$$

It follows that under this model, the dominance coefficient and effect size for the contribution to disease risk equal those for fitness (justifying our use of the same notation for the $\alpha$'s in both).

We now return to the contribution of individual loci to disease risk under this model. Assuming that each locus has a small contribution, i.e., that $\alpha_j(g) \ll 1$ (which follows from $s_j \ll 1$) for $j = 1, \ldots, L$, we can approximate the variance in disease risk by

$$V(P) \approx \exp(-2\sum_{j=1}^{L} E(\alpha_j(g_j)))\sum_{j=1}^{L} V(\alpha_j(g_j)). \tag{A1.6}$$

In other words, the contribution of an individual locus to variation in disease risk is proportional to the variance in fitness at that locus. Here, we consider semi-dominant and recessive loci for which the variances are

$$V(x; s, \frac{1}{2}) = \frac{1}{2}s^2 x(1-x) \tag{A1.7}$$

and

$$V(x; s, 0) = s^2 x^2 (1 - x^2), \tag{A1.8}$$

correspondingly.

### 4.2 Demographic effects on the variance

Supplementary Figure A1.5 depicts how different allele frequencies at semi-dominant and recessive loci contribute to the variance in disease risk under the Tennessen et al. [2] model (expanding on Figure 4 in Chapter 1). Because we consider only one selection coefficient at a time, the relationship between effect sizes and selection coefficient has no effect here; however, we do assume that the dominance coefficient for fitness and for disease risk are the same. The graphs can also be interpreted as the proportional contribution of different allele frequencies to the variance in fitness

among individuals. To elucidate the effects of recent demographic events, we also show results for the model with a constant population size (equivalent to the one for the African population before the onset of growth) and for a population that experienced the same instantaneous increase in population size as the ancestral African population in the Tennessen et al. model but then remained constant (from $\sim 7,000$ to $\sim 14,500$ around $6,000$ generations ago, cf. Supplementary Figure A1.1A), which we refer to as the "older growth" model.

**Demographic effects in the semi-dominant case.** First, we consider the effectively neutral regime (Supplementary Figure A1.5A). In the model with constant population size, the proportional contribution is uniform across frequencies, as expected [3]. In the model of older growth, there is an increased contribution of low and high frequency alleles to the variance (as diversity patterns did not have sufficient time to reach equilibrium yet). In the model for Africans, a similar pattern is observed, with a tiny increase in the contribution from rare alleles due to recent growth (amounting to $0.41\%$ of variance in deleterious variants with frequency below $0.1\%$ and $0.4\%$ in variants above $99.9\%$). In the model for Europeans, the increase due to growth is also negligible ($0.61\%$ of variance in variants with frequency below $0.1\%$ and $0.6\%$ in variants above $99.9\%$). However, the bottleneck leads to an increased contribution of intermediate frequencies at the expense of moderately low and high frequency alleles (since low and high frequency alleles are quickly lost or fixed after the reduction in population size).

In the weak selection regime (Supplementary Figure A1.5B), selection leads to a shift towards lower frequencies and thus to an increased contribution to variance of lower frequency alleles. In turn, the effect of older growth is to increase the contribution of high frequencies: the reason being that before the increase in population size, a greater proportion of sites is fixed for the deleterious allele and at such sites, "normal" mutations lead to high frequency deleterious alleles. The recent growth in the model for Africans further causes a small increase in the contribution of rare alleles (amounting to $1.4\%$ of variance in variants with frequency below $0.1\%$ and $0.07\%$ in variants above $99.9\%$). In the model for Europeans, this increase is also small ($1.9\%$ of variance in variants with frequency below $0.1\%$ and $0.1\%$ in variants above $99.9\%$), but the bottleneck again has a substantial effect, increasing the contribution of intermediate frequencies at the expense of lower and higher frequencies.

In the strong selection regime, because of the quick turnover of deleterious alleles, the older increase in population size and the bottleneck in Europeans are too far in the past to have had an effect on alleles that are currently segregating (Supplementary Figure A1.5C). By the same token, in the Tennessen et al. model, alleles segregating at present are young and therefore the recent growth

resulted in a decrease in their frequencies (cf. section 2.3), substantially increasing the contribution of rare alleles to variance (with $\sim 70\%$ of the variance contributed by alleles at frequency below $0.1\%$).

**Demographic effects in the recessive case.** In this case, recent growth has little effect in all selection regimes. The contribution of low frequency alleles to variance is much smaller because their effect on load or disease risk is manifested only in homozygotes (Supplementary Figure A1.5D-F). As a result, the increase in the number of rare deleterious alleles caused by recent growth has a negligible effect on their contribution to the variance in disease risk under both the model for Europeans and Africans (amounting to $\sim 10^{-4}\%$ in the neutral regime, $\sim 5 \cdot 10^{-4}\%$ in the weakly selected and $\sim 0.01\%$ in the strongly selected regime, in variants with frequency below $0.1\%$). In turn, the increase in the number of high frequency alleles (due to "normal" mutants on a deleterious background) has a higher impact but it is still quite small (amounting to $\sim 1\%$ in the neutral regime and $\sim 0.2\%$ in the weakly selected regime that are due to variants with frequency above $99.9\%$).

In the weak and strong selection regimes, there is a peak in the contribution to variance at intermediate frequency (Supplementary Figure A1.5E and F). Moving from low to intermediate frequencies, the contribution to the variance of a mutant allele increases (see Equation A1.8). This increase is halted, however, because at higher frequencies, selection on homozygotes for the deleterious allele kicks in, leading to few alleles at high frequencies. (Specifically, for a constant population size and given a low mutation rate, the frequency spectrum of deleterious alleles is well approximated by $C\frac{e^{-\alpha x^2}}{x}$, where $C$ is a normalizing constant [3], and thus the contribution to variance can be approximated by $De^{-\alpha x^2}x(1-x)^2$, where $D$ is a normalizing constant.) In the model for Africans (and for older growth), this peak is at higher frequencies in the weak selection regime (Supplementary Figure A1.5E), because the older increase in population size led to relatively more high frequency alleles at present.

The bottleneck in the model for Europeans has a much more pronounced effect, causing a shift toward intermediate allele frequencies and a corresponding shift in the contribution to variance in all selection regimes (Supplementary Figure A1.5D-F). As opposed to the semi-dominant case, this is also true for the strong selection regime, as recessive deleterious alleles can reach substantial allele frequencies.

**Summary.** Population growth increases the relative proportion of rare alleles and could therefore be expected to increase their relative contribution to the variance in disease risk. However, because rare alleles contribute less to the variance to begin with, this effect may be relatively small. Assess-

ing the effects of growth on the genetic architecture of disease risk therefore requires quantification. Here, we have shown that, at least based on current estimates of recent growth, the effects on the variance in disease risk are expected to be negligible. The one exception is the case of strongly selected quasi-dominant alleles, which are young and therefore whose frequencies do reflect the recent population size expansion. Interestingly, in this case, while the architecture of disease risk is substantially affected by growth, the expected load (or disease prevalence) remains unchanged, i.e., the same load will be due to many more deleterious alleles that segregate at lower frequencies than had the population not grown.

In contrast to growth, the bottleneck in European populations should have increased the proportion of intermediate frequency deleterious alleles at the expense of low and high frequency ones (with the exception of strongly selected quasi-dominant alleles, because they are so young). In other words, in these populations, there will be only a small effect on load but a substantial effect on the architecture of disease, with a greater proportion of the variance in disease risk due to intermediate frequency alleles.

### 4.3 The contribution of rare alleles in a mixture model

In reality, we expect that the variants underlying a complex disease will have a variety of selection coefficients and effect sizes rather than a single one. Under a model with such a mixture, the expected contributions of different allele frequencies to the variance in disease risk can be derived as follows. For simplicity, assume that mutations are semi-dominant (so the dominance coefficient is dropped from the notation). At a site with selection coefficient $s$, the expected contribution to the variance from deleterious alleles below frequency $\omega$ is

$$V_\omega(s) = \frac{1}{2} C E(a^2|s) \int_0^\omega f(x;s)x(1-x)dx, \tag{A1.9}$$

where $E(a^2|s)$ is the expectation of the effect size squared for sites with selection coefficient $s$, $f(x;s)$ is the probability of the deleterious allele being at frequency $x$ (here, we do not condition of the allele segregating) and the proportion coefficient $C$ is akin to the first term in Equation A1.5. The overall contribution to variance of a site is $V_1(s)$ and the fraction of that contribution coming from variants below frequency $\omega$ is $\Theta_\omega(s) \equiv \frac{V_\omega(s)}{V_1(s)}$. When all sites are considered jointly, denoting the input of mutations with selection coefficient $s$ by $\mu(s)$, the expected proportion of variance from deleterious alleles below frequency $\omega$ is then

$$\Theta_\omega = \frac{\int_s \mu(s) V_1(s) \Theta_\omega(s) ds}{\int_s \mu(s) V_1(s) ds}. \tag{A1.10}$$

Examining the terms in Equation A1.10 suggests that the contribution of rare alleles depends strongly on the relationship between effect sizes and selection coefficients. Specifically, the proportional contribution of rare alleles $\Theta_{0.1\%}(s)$ becomes substantial only for strong selection coefficients (Figure 4D in Chapter 1), as shown in section 4.2. The behavior of the overall contribution to variance $V_1(s)$, however, depends on the relationship between effect sizes and selection coefficients. If we assume that the effect sizes do not depend on the selection coefficients (or more precisely that $E(a^2|s)$ is constant) then $V_1(s)$ from weakly selected sites is much greater than from strongly selected sites (Figure 4E in Chapter 1) and rare alleles will make an important contribution only if a very large fraction of the mutational input is at strongly selected sites. If we assume the other extreme in which the effect sizes are proportional to the selection coefficient (or more precisely that $E(a^2|s) \propto s^2$, as in the model in section 4.1) then $V_1(s)$ strongly increases with the $s$ (Figure 4E in Chapter 1) and rare alleles would make an important contribution unless the fraction of the mutational input at strongly selected sites is very small. In reality, the outcome could be anywhere in between.

As an illustration, we consider a simple model in which we vary the correlation between selection on variants and their effect on a trait. We assume that half of the newly arising mutations have a weak selection coefficient $s_w = 0.0002$ and half have a strong selection coefficient of $s_s = 0.01$. For strongly selected mutations, the effect size on the trait, $a$, is chosen to be $cs_s$ with probability $\frac{1}{2}(1+p)$ and $cs_w$ with probability $\frac{1}{2}(1-p)$, where $c$ is a positive constant and $0 \leq p \leq 1$; correspondingly, for weakly selected mutations the effect size is chosen to be $cs_w$ with probability $\frac{1}{2}(1+p)$ and $cs_s$ with probability $\frac{1}{2}(1-p)$. In this model, the marginal distributions of selection coefficients and effect sizes do not depend on $p$, while the correlation between them is equal to $p$. To obtain Figure 4F in Chapter 1 we therefore vary $p$ between 0 and 1.

# 5 Tables and Figures

**Supplementary table A1.1:**
**Changes to load under the bottleneck and growth models**

| | | | Effectively neutral | Weak | | Strong |
|---|---|---|---|---|---|---|
| | | | | closer to neutral | closer to strong | |
| Bottleneck | Semi-dominant | fixed | increase | increase | increase | — |
| | | segregating | decrease | decrease | increase | unchanged |
| | | total | unchanged | increase | increase | unchanged |
| | Recessive | fixed | increase | increase | increase | — |
| | | segregating | decrease | decrease | increase | transient increase |
| | | total | unchanged | increase | increase | transient increase |
| Growth | Semi-dominant | fixed | decrease | decrease | | — |
| | | segregating | increase | increase | | unchanged |
| | | total | unchanged | unchanged | | unchanged |
| | Recessive | fixed | decrease | decrease | | — |
| | | segregating | increase | increase | | transient decrease |
| | | total | unchanged | unchanged | | transient decrease |

Supplementary Table A1.1: Changes to load under the bottleneck and growth models. The effects on fixed, segregating and total load are depicted by selection regime. The symbol — denotes the cases in which there is no contribution to load both before and after the change in population size.

**Supplementary table A1.2:**

**Estimated mean frequencies in AAs and EAs at different classes of sites**

| Method | Chr. | Category | # SNVs | $AA_{Mean}$ | $AA_{SE}$ | $EA_{Mean}$ | $EA_{SE}$ | t-score |
|---|---|---|---|---|---|---|---|---|
| Non-coding | Aut | — | 300209 | 0.034 | 0.00026 | 0.034 | 0.00028 | 0.44 |
| Non-coding | X | — | 8355 | 0.030 | 0.0015 | 0.028 | 0.0016 | 1.1 |
| Synonymous | Aut | — | 220391 | 0.033 | 0.00030 | 0.033 | 0.00032 | 0.87 |
| Synonymous | X | — | 7001 | 0.028 | 0.0016 | 0.029 | 0.0018 | -0.10 |
| Non-synonymous | Aut | — | 351265 | 0.014 | 0.00015 | 0.014 | 0.00016 | 0.40 |
| Non-synonymous | X | — | 10293 | 0.012 | 0.00086 | 0.012 | 0.00095 | 0.076 |
| PolyPhen2 | Aut | D | 121280 | 0.0078 | 0.00011 | 0.0076 | 0.00012 | 1.2 |
| PolyPhen2 | Aut | P | 65400 | 0.012 | 0.00018 | 0.012 | 0.00020 | 0.52 |
| PolyPhen2 | Aut | B | 132047 | 0.019 | 0.00024 | 0.019 | 0.00026 | 0.55 |
| PolyPhen2 | X | D | 3205 | 0.0072 | 0.00065 | 0.0079 | 0.00078 | -0.99 |
| PolyPhen2 | X | P | 1957 | 0.013 | 0.0012 | 0.012 | 0.0012 | 0.98 |
| PolyPhen2 | X | B | 3948 | 0.014 | 0.0011 | 0.014 | 0.0012 | 0.044 |
| Sift | Aut | D | 145986 | 0.0095 | 0.00012 | 0.0093 | 0.00013 | 1.6 |
| Sift | Aut | T | 180091 | 0.018 | 0.00021 | 0.018 | 0.00022 | -0.13 |
| Sift | X | D | 4251 | 0.0099 | 0.00076 | 0.0096 | 0.00082 | 0.34 |
| Sift | X | T | 5517 | 0.017 | 0.0013 | 0.017 | 0.0015 | -0.29 |
| LRT | Aut | D | 146701 | 0.0060 | 8.5e-05 | 0.0060 | 9.5e-05 | -0.11 |
| LRT | Aut | N | 160179 | 0.020 | 0.00024 | 0.020 | 0.00026 | 0.20 |
| LRT | Aut | U | 13845 | 0.0066 | 0.00036 | 0.006 | 0.00039 | 2.6 |
| LRT | X | D | 3270 | 0.0038 | 0.00037 | 0.0034 | 0.00034 | 0.93 |
| LRT | X | N | 4548 | 0.017 | 0.0014 | 0.017 | 0.0016 | -0.37 |
| LRT | X | U | 886 | 0.0052 | 0.0013 | 0.0046 | 0.0015 | 0.40 |
| MutationTaster | Aut | D | 155138 | 0.0022 | 2.9e-05 | 0.0017 | 3.0e-05 | 18 |
| MutationTaster | Aut | A | 5089 | 0.00089 | 9.5e-05 | 0.00056 | 4.8e-05 | 4.3 |
| MutationTaster | Aut | N | 161169 | 0.0062 | 6.8e-05 | 0.0047 | 6.7e-05 | 21 |
| MutationTaster | Aut | P | 9040 | 0.36 | 0.0047 | 0.39 | 0.0051 | -6.5 |
| MutationTaster | X | D | 3860 | 0.021 | 0.0021 | 0.023 | 0.0023 | -1.2 |
| MutationTaster | X | A | 76 | 0.0010 | 0.00058 | 0.00039 | 0.00017 | 1.5 |
| MutationTaster | X | N | 5566 | 0.0030 | 0.00026 | 0.0013 | 0.00022 | 7.0 |
| MutationTaster | X | P | 131 | 0.16 | 0.028 | 0.16 | 0.029 | 0.28 |

Supplementary Table A1.2: Comparison of mean frequencies in AAs and EAs at different classes of sites, classified according to whether the sites are on the autosomes or X, and using a variety of different functional classifications (after application of our bias-correction method). For this table, the data were subsampled down to 3852 chromosomes for AAs and EAs each, to enable X vs autosome comparisons. Note that the mean frequencies in each row are not significantly different ($|t - score| < 2$, with the sole exception of the functional classifications from MutationTaster (which are highly significant). The unusual results for MutationTaster likely arise because MutationTaster uses previously estimated population frequencies in its classification, thus introducing further biases for population genetic analysis that are not properly addressed by correction method.

**Supplementary table A1.3:**
**Estimated mean frequencies with and without bias correction**

| Method | Chr. | Category | Without bias correction | | | | With bias correction | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $AA_{Mean}$ | $AA_{SE}$ | $EA_{Mean}$ | $EA_{SE}$ | $AA_{Mean}$ | $AA_{SE}$ | $EA_{Mean}$ | $EA_{SE}$ |
| Non-synonymous | Aut | — | 0.014 | 0.00015 | 0.014 | 0.000162 | 0.014 | 0.00015 | 0.014 | 0.00016 |
| PolyPhen2 | Aut | D | 0.0038 | 9.3E-05 | 0.0033 | 1.0E-04 | 0.0078 | 0.00011 | 0.0076 | 0.00012 |
| PolyPhen2 | Aut | P | 0.0060 | 0.00017 | 0.0053 | 0.00019 | 0.012 | 0.00018 | 0.012 | 0.00020 |
| PolyPhen2 | Aut | B | 0.026 | 0.00035 | 0.026 | 0.00037 | 0.019 | 0.00024 | 0.019 | 0.00026 |
| Sift | Aut | D | 0.0061 | 0.00013 | 0.0055 | 0.00014 | 0.0095 | 0.00012 | 0.0093 | 0.00013 |
| Sift | Aut | T | 0.020 | 0.00026 | 0.021 | 0.00028 | 0.018 | 0.00021 | 0.018 | 0.00022 |
| LRT | Aut | D | 0.0028 | 6.4E-05 | 0.0025 | 7.4E-05 | 0.0060 | 8.5e-05 | 0.0060 | 9.5e-05 |
| LRT | Aut | N | 0.023 | 0.00029 | 0.023 | 0.00031 | 0.020 | 0.00024 | 0.020 | 0.00026 |
| LRT | Aut | U | 0.0081 | 0.00048 | 0.0071 | 5.0E-04 | 0.0066 | 0.00036 | 0.006 | 0.00039 |
| MutationTaster | Aut | D | 0.0017 | 4.3E-05 | 0.0011 | 4.3E-05 | 0.0022 | 2.9e-05 | 0.0017 | 3.0e-05 |
| MutationTaster | Aut | A | 0.0013 | 0.00034 | 0.00099 | 0.00032 | 0.00089 | 9.5e-05 | 0.00056 | 4.8e-05 |
| MutationTaster | Aut | N | 0.013 | 0.00024 | 0.012 | 0.00025 | 0.0062 | 6.8e-05 | 0.0047 | 6.7e-05 |
| MutationTaster | Aut | P | 0.26 | 0.0027 | 0.30 | 0.0032 | 0.36 | 0.0047 | 0.39 | 0.0051 |

Supplementary Table A1.3: Comparison of estimated mean frequencies in samples of 3852 chromosomes, with and without bias correction of the functional annotations. Recall that we observed that all four functional prediction methods typically have low probabilities of assigned 'damaging' status to SNVs where the genome reference carries the derived allele. Notice that prior to applying the bias correction (using all SNVs), AAs tend to have higher allele frequencies at putatively damaging sites, as reported by Tennessen et al. This is likely because most of the reference genome is of non-African origin. After applying our bias correction, we observe that AAs and EAs have essentially identical allele frequencies in all functional categories (except for MutationTaster, likely for reasons discussed above).

**Supplementary table A1.4:**
**Estimated mean frequencies using different methods for classifying sites**

| Category | $AA_{Mean}$ | $AA_{SE}$ | $EA_{Mean}$ | $EA_{SE}$ | T-Stat |
|---|---|---|---|---|---|
| Uncorrected (biased) PolyPhen Scores | | | | | |
| Prob. Damaging | 0.00277 | 6.79e-05 | 0.00239 | 7.31e-05 | 5.4 |
| Poss. Damaging | 0.00452 | 0.00013 | 0.00401 | 0.00014 | 3.84 |
| Benign | 0.0208 | 0.000278 | 0.0212 | 0.000297 | -1.34 |
| Bias-corrected PolyPhen Scores | | | | | |
| Prob. Damaging | 0.00593 | 8.11e-05 | 0.00582 | 8.76e-05 | 1.23 |
| Poss. Damaging | 0.00955 | 0.00014 | 0.00948 | 0.000151 | 0.488 |
| Benign | 0.0154 | 0.000186 | 0.0153 | 2e-04 | 0.527 |
| Human-independent PolyPhen Scores | | | | | |
| 3<PSIC | 0.0056 | 0.0002 | 0.0054 | 0.0003 | 0.45 |
| 1.5<PSIC<3 | 0.011 | 0.0002 | 0.011 | 0.0002 | -0.06 |
| PSIC<1.5 | 0.019 | 0.0003 | 0.019 | 0.0003 | -0.07 |

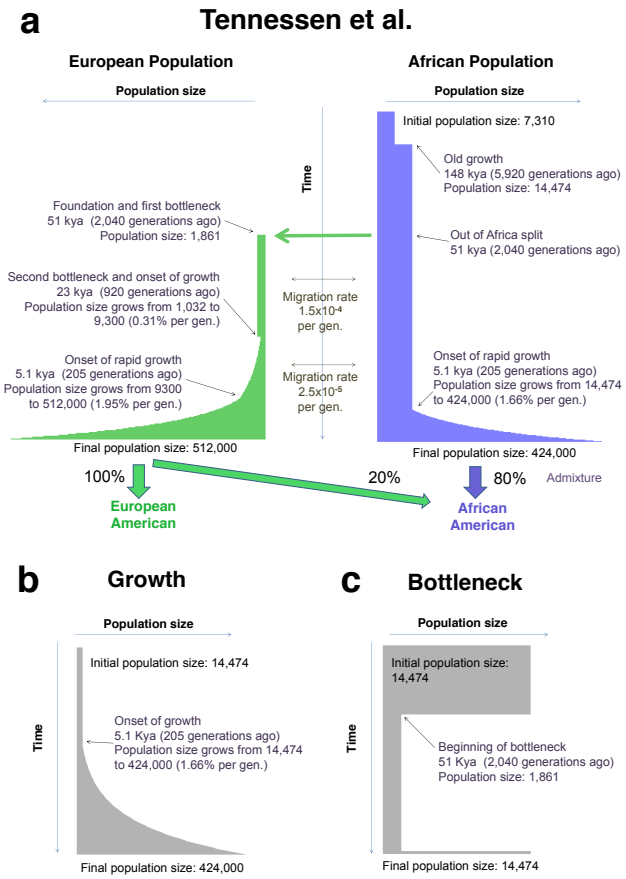
Supplementary Table A1.4: Comparison of estimated mean frequencies at autosomal nonsynonymous sites in the Fu et al data, using the full autosomal samples. The top block of data use the uncorrected (biased) PolyPhen scores, and suggest significant differences between populations. The middle block of data applies our bias correction, and shows no significant differences between populations. The bottom block of data uses an unpublished version of the PolyPhen "PSIC" scores that are calculated independent of the human reference sequence, and hence are unbiased (kindly provided by the Shamil Sunyaev lab). These too show no significant difference between populations. Note that DAFs differ between the second two blocks of data due to arbitrary choices in score cutoffs.

**Supplementary table A1.5:**
**Summary of 1000 Genomes analysis**

| Category | $YRI_{Mean}$ | $YRI_{SE}$ | $CEU_{Mean}$ | $CEU_{SE}$ | P-value |
|---|---|---|---|---|---|
| Individual-Level Counts | | | | | |
| Synonymous | 18,141 | 119 | 17,992 | 122 | N.S. |
| Nonsynonymous | 9903 | 104 | 9825 | 80 | N.S. |
| Prob. Damaging | 2153 | 31 | 2111 | 26 | N.S. |
| Poss. Damaging | 1851 | 27 | 1836 | 24 | N.S. |
| Benign | 5899 | 67 | 5878 | 55 | N.S. |

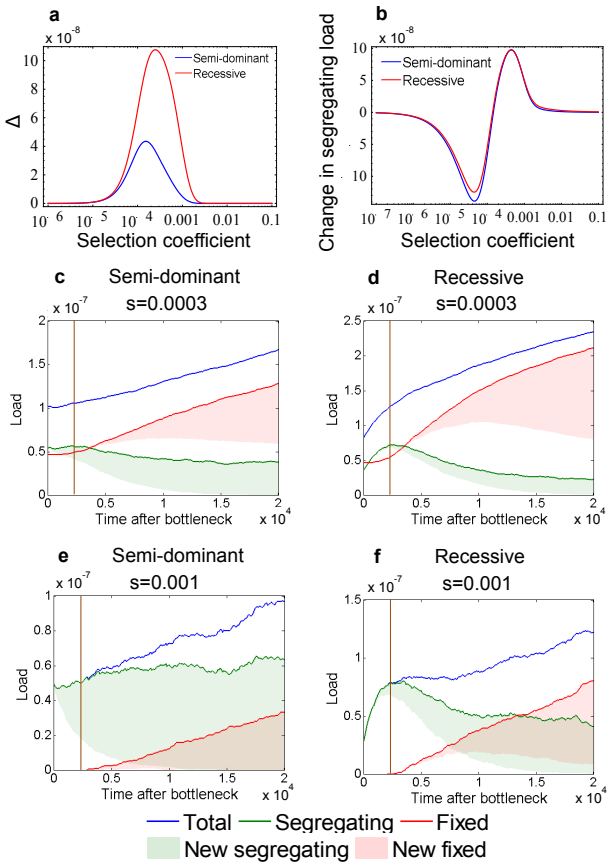
Supplementary Table A1.5: Summary of 1000 Genomes Analysis. This table shows the mean numbers of derived alleles per individual in the YRI and CEU populations. The functional categories (Probably/Possibly Damaging and Benign) were obtained from PolyPhen, and adjusted using our bias correction method. SEs obtained by bootstrapping across SNVs. We also obtained identical conclusions (i.e., no difference between populations) when the analysis was done in terms of DAFs, and also when we used the human-independent PolyPhen (PSIC) scores.

**Supplementary figure A1.1:**
**Demographic scenarios**




Supplementary Fig. A1.1: The three demographic models that we consider. A) The Out-of-Africa model estimated by Tennessen et al. [2]. C) Exponential growth.B) A population bottleneck. All population sizes are given as number of diploid individuals. In some cases, in order to study the equilibration process, we extend the growth scenario to include a priod with a constant population size after growth and the bottleneck model to include a longer period with a reduced population size.

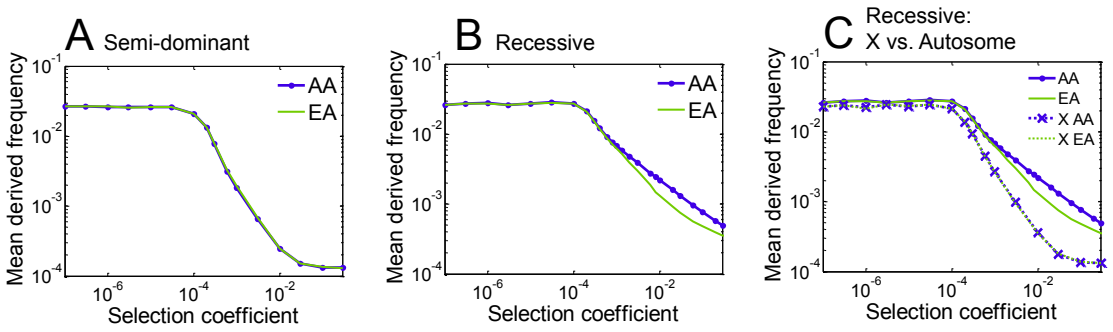
**Supplementary figure A1.2:**
**Changes in load shortly after a bottleneck**




Supplementary Fig. A1.2: The changes in load shortly after a bottleneck. The figure shows (A) the expected change in fixed load due to mutations that segregated before the bottleneck and (B) the expected change in segregating load due to the bottleneck as a function of the selection coefficient. Shown are segregating, fixed and total load from new and all mutations as a function of time since the population size decrease. The semi-dominant (C and E) and recessive cases (D and F) are shown with a selection coefficient in the weak selection regime closer to neutral ($s = 0.0003$) and closer to strong ($s = 0.001$).

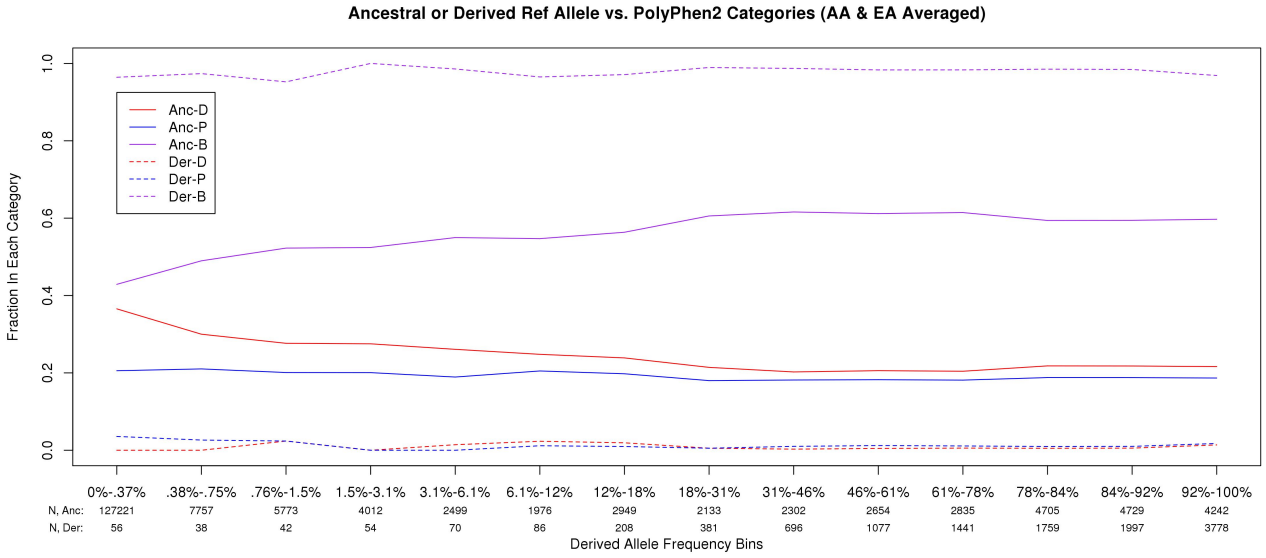
**Supplementary figure A1.3:**

**Predicted mean derived frequencies as a function of selection coefficient**




Supplementary Fig. A1.3: Mean derived frequencies predicted as a function of selection coefficient, for the AA and EA demographies. Notice that in (**A**) we predict that for semi-dominant sites AAs and EAs should have essentially identical mean derived frequencies for all levels of selection. In (**B**) we predict a small increase in mean frequencies for AAs at recessive sites with moderate-strong selection. (**C**) provides X vs autosome comparisons under the recessive model; note that recessive alleles on the X experience selection as dominant alleles in males.

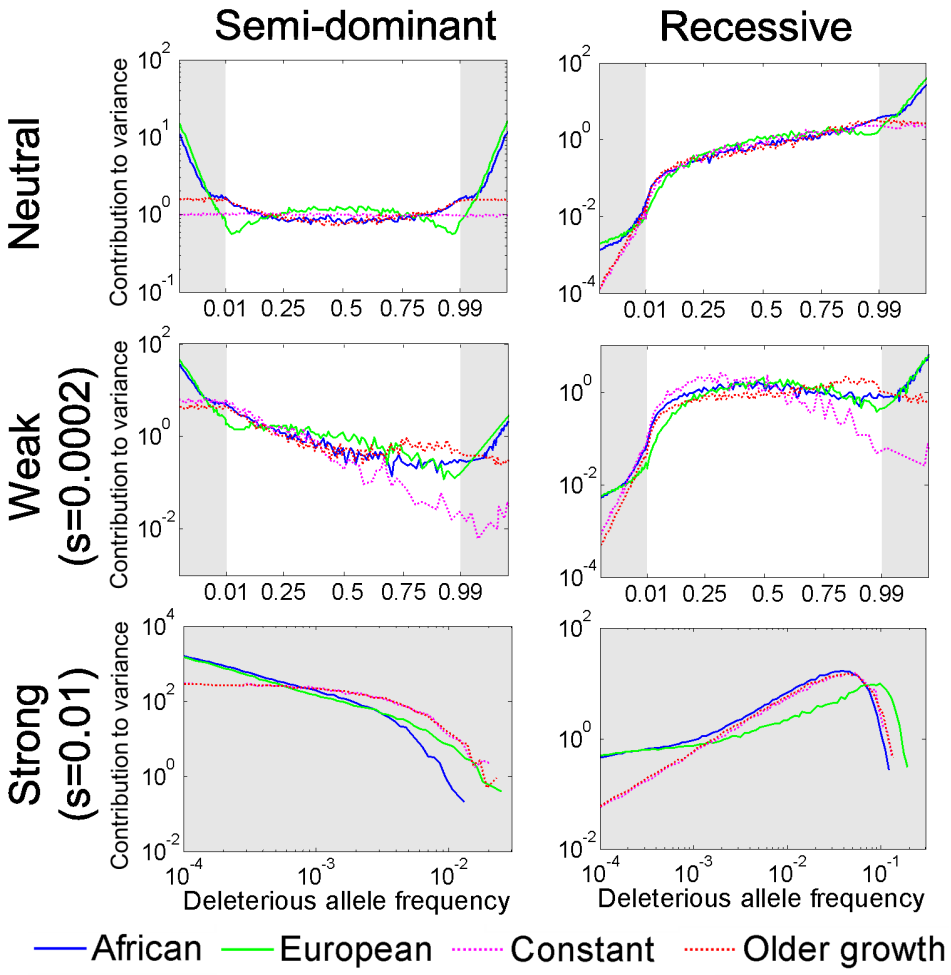
**Supplementary figure A1.4:**
**Reference bias in PolyPhen 2**




Supplementary Fig. A1.4: Illustration of the reference bias present in PolyPhen 2 [22]. The other functional prediction methods that we considered have a similar bias. The x-axis shows the mean population frequency of nonsynonymous SNVs in the Fu et al data (the left-most bins cover very narrow intervals of frequencies since most of the data are present in these bins). The y-axis plots the fraction of SNVs in each bin that are classified into each of the three PolyPhen categories: **B**enign, **P**ossibly damaging, Probably **D**amaging; and shown separately according to whether the genome reference sequence carries the ancestral or the derived allele. Notice that when the reference carries the ancestral allele, an SNV is classified as Damaging with a probability that ranges from nearly 40% at low frequencies to ≈20% at high frequencies (solid red line). In contrast, for SNVs where the reference carries the derived allele, the fraction of Damaging alleles is near 0% at all frequencies (dotted red line).

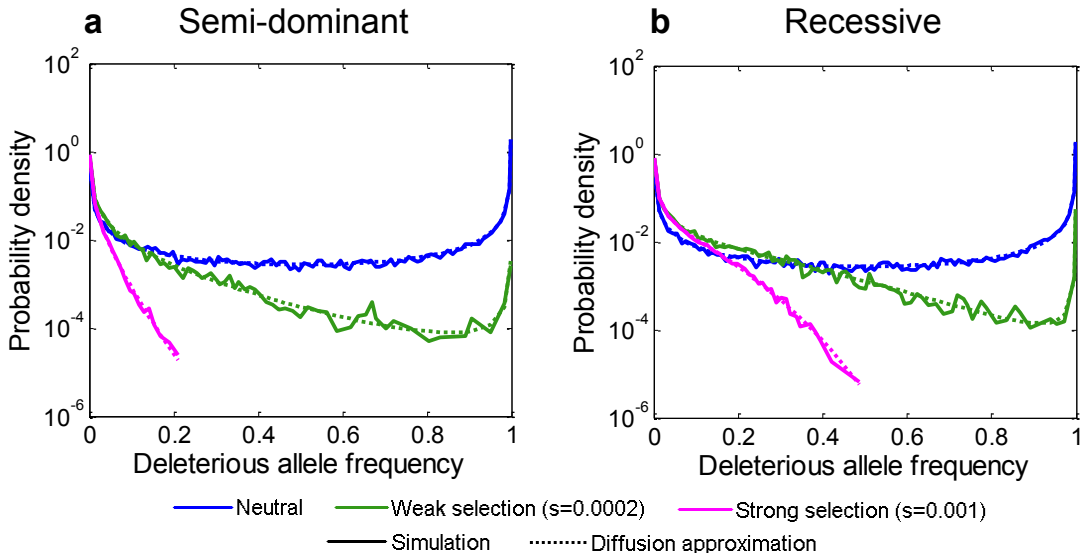
**Supplementary figure A1.5:**

**Contribution of different allele frequencies to variance in disease risk**




Supplementary Fig. A1.5: The proportional contribution of different allele frequencies to variance in disease risk, under the Tennessen et al. model for Africans and Europeans. Shaded regions correspond to a logarithmic scale on the x-axis, which is included to show the (minor) effects of recent growth.

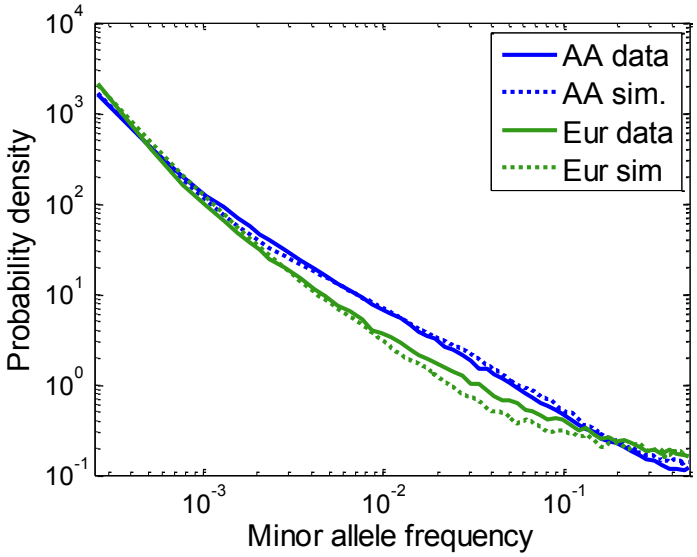
**Supplementary figure A1.6:**
**Comparison of theoretical and simulated frequency spectra**




Supplementary Fig. A1.6: Comparison of theoretical and simulated frequency spectra for a constant population size in the (A) semi-dominant and (B) recessive models. Shown are the results based on the diffusion approximation (solid) and on simulations (dashed) for several selection coefficients. The population size was taken as $N = 14,474$ and the mutation rate as $u = 2.36 \cdot 10^{-8}$ per generation per site. The number of runs for each set of parameters was $10^6$.

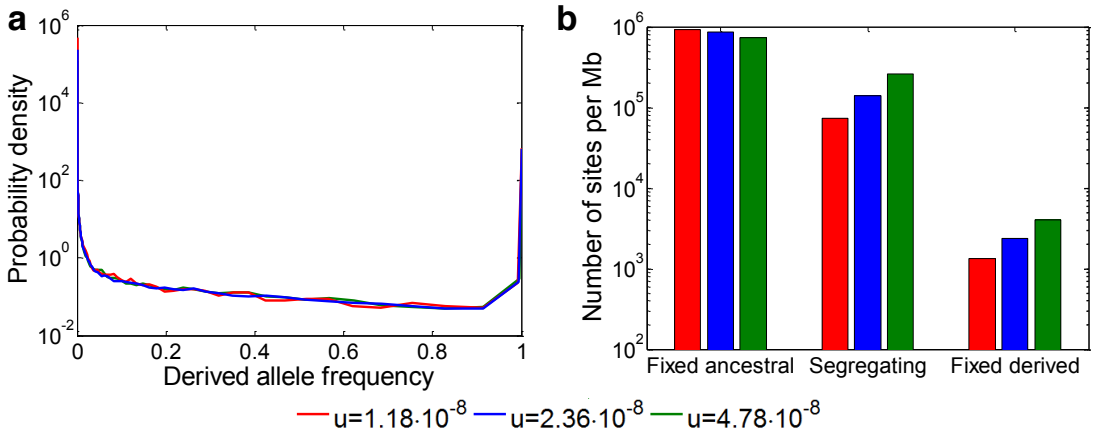
**Supplementary figure A1.7:**

**Comparison of estimated and simulated frequency spectra**



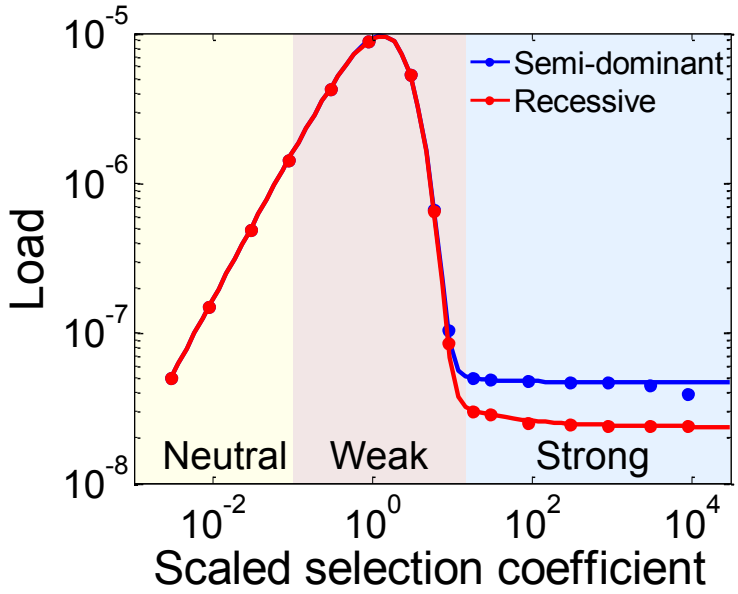

Supplementary Fig. A1.7: Comparison of the minor allele frequency spectrum in data from Fu et. al. and in simulations based on the Tennessen et al. model. The spectra are for a sample size of 3852 chromosomes in AA and EA populations, for both the data and simulations.

**Supplementary figure A1.8:**
**Sensitivity to mutation rate**




Supplementary Fig. A1.8: Sensitivity of (A) the frequency spectrum and (B) the number of segregating and fixed sites to the mutation rate. The results are shown for simulations of the African population but are qualitatively similar for the European population.

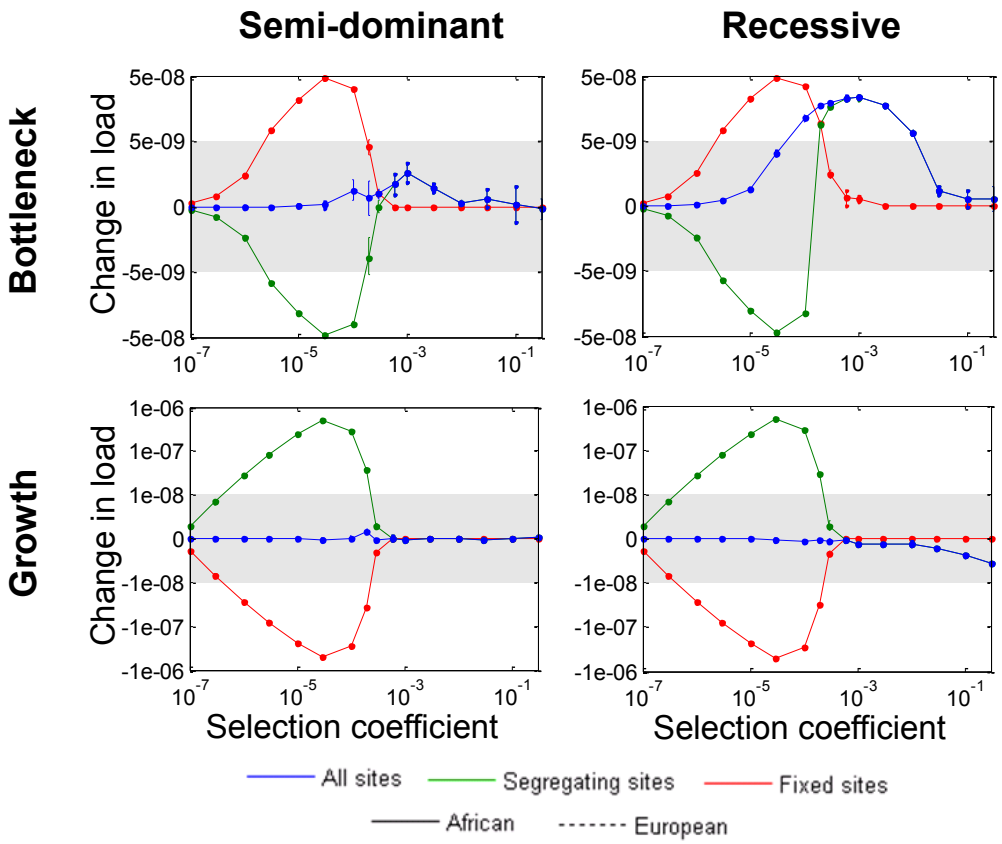
**Supplementary figure A1.9:**
**Load in a population of constant size**




Supplementary Fig. A1.9: Load as a function of selection coefficient in a population of constant size. Results are shown for the semi-dominant (blue) and recessive models (red), where the diffusion approximation is shown as a solid line and simulation results as circles. The population size is $N = 14,474$.

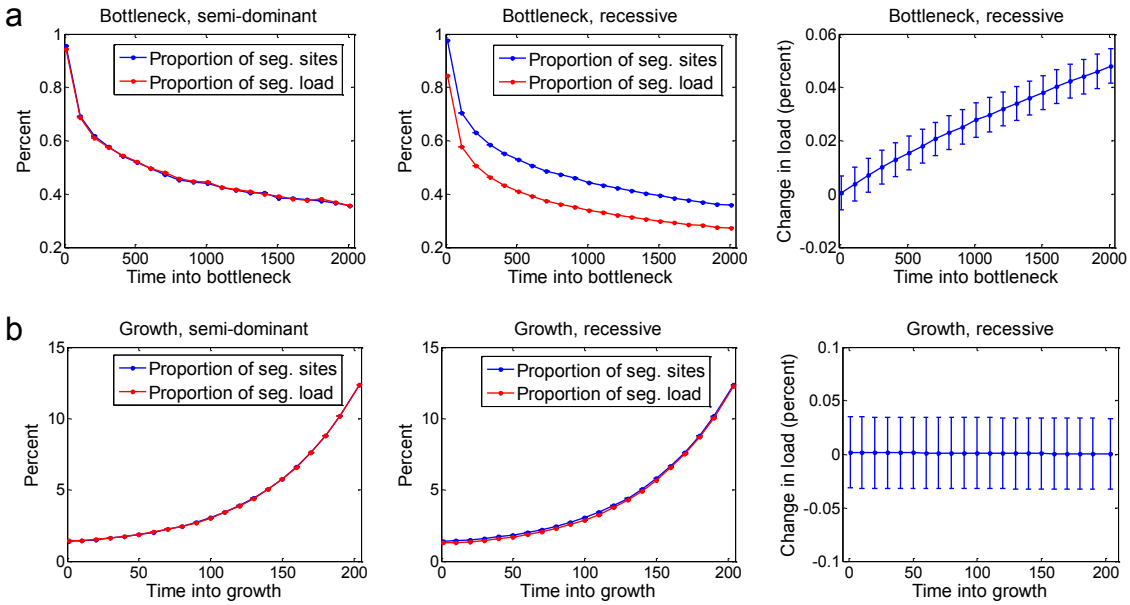
**Supplementary figure A1.10:**

**Changes to load under the bottleneck and growth models**




Supplementary Fig. A1.10: The changes to the segregating, fixed and total load under the bottleneck and growth models. Analogous graphs for the Tennessen et al. model are presented in Figure 3 of Chapter 1. Changes are measured by comparison to a population in which the population size has remained constant at the size that it was at the beginning of the demographic model. In the shaded areas, load is shown on linear scale; otherwise it is shown on logarithmic scale.

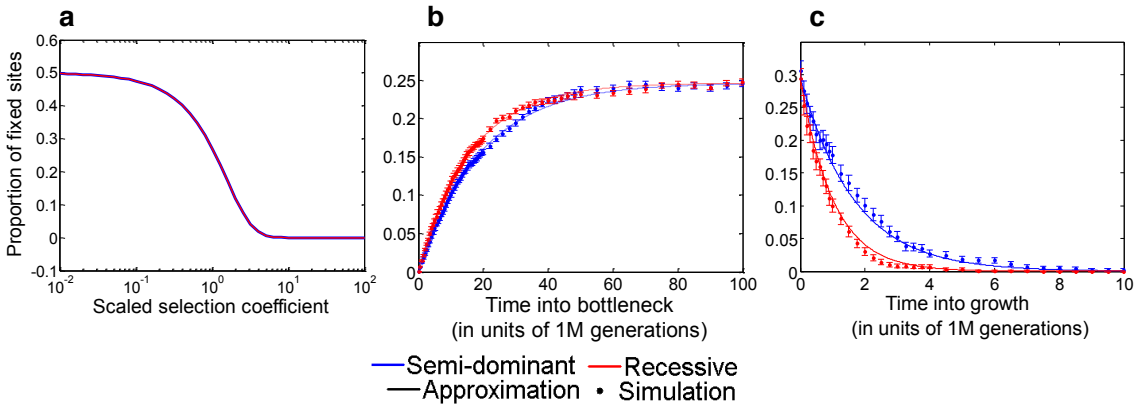
**Supplementary figure A1.11:**
**Load in the effectively neutral regime**




Supplementary Fig. A1.11: Segregating and total load in the bottleneck and growth models in the effectively neutral regime. The proportion of segregating sites, their proportional contribution to load, and the proportional change in total load are shown as a function of time (A) after the bottleneck and (B) since the onset of growth. The selection coefficient is $s = 10^{-7}$. In the semi-dominant case, the expected total load is always $s/2$ regardless of changes in population size; in the recessive case, changes to the proportion of segregating sites affect the total load, but this effect is negligibly small.

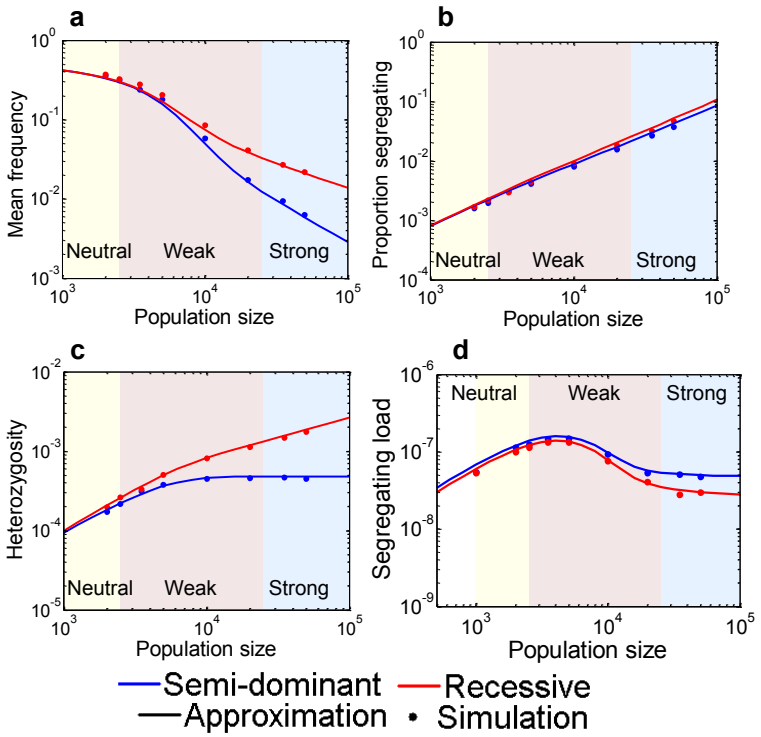
**Supplementary figure A1.12:**

**Fixed sites in the weak selection regime**




Supplementary Fig. A1.12: Proportion of sites fixed for deleterious alleles in the weak selection regime. In all graphs, the selection coefficient is $s = 10^{-4}$. (A) The equilibrium proportion as a function of the scaled selection coefficient ($\alpha = 2Ns$), where the population size was varied. (B) The proportion as a function of time after the change in population size in the bottleneck model. (C) The proportion as a function of time after the change in population size in the growth model.

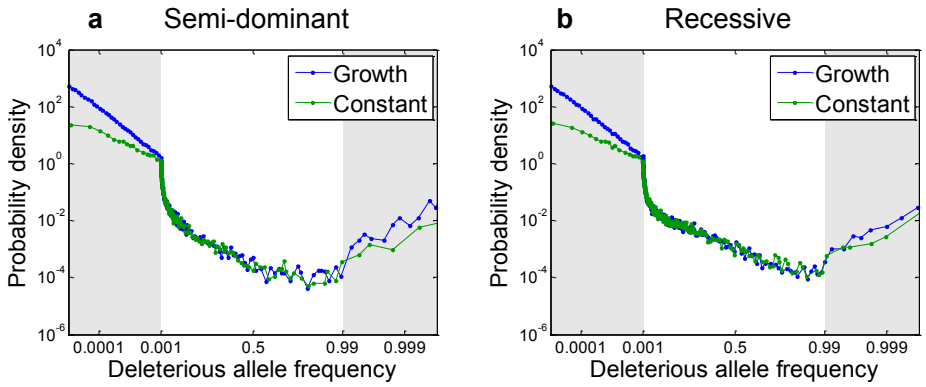
**Supplementary figure A1.13:**
**Equilibrium properties of segregating sites**




Supplementary Fig. A1.13: Equilibrium properties of segregating sites as a function of population size in constant population size models. In all graphs, $s = 2 \cdot 10^{-4}$. (A) The average frequency of segregating deleterious alleles. (B) The proportion of segregating sites. (C) Heterozygosity. (D) Segregating load.

**Supplementary figure A1.14:**

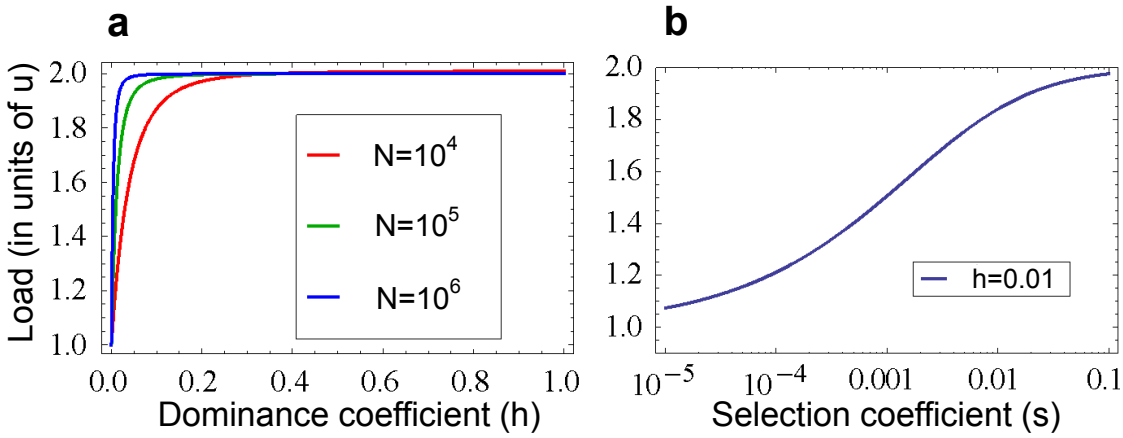
**Frequency spectrum of weakly deleterious sites with and without growth**




Supplementary Fig. A1.14: The frequency spectrum of weakly deleterious segregating sites in models with and without growth. In the shaded areas, frequency is shown on logarithmic scale; otherwise it is shown on linear scale.

**Supplementary figure A1.15:**

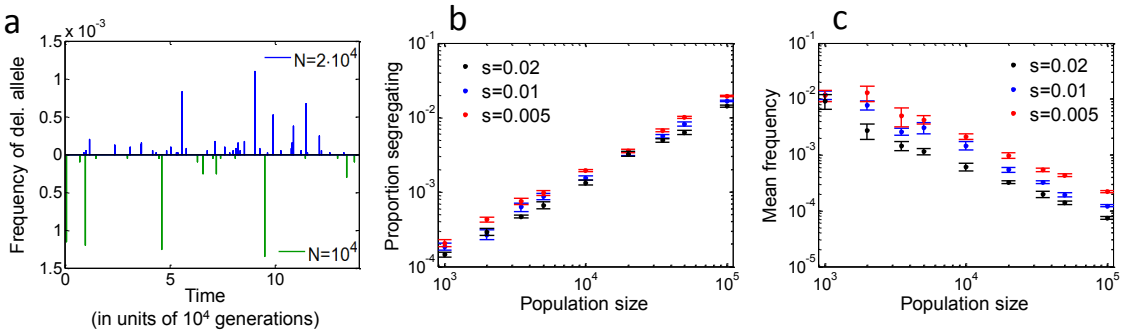
**Dependence of the load on the dominance coefficient at equilibrium**




Supplementary Fig. A1.15: The dependence of the load on the dominance coefficient at equilibrium. The graphs were generated using the diffusion approximation for the stationary distribution assuming that the deleterious allele frequency is small [3]. A) Load as a function of the dominance coefficient $h$, with $s = 0.01$ and population size $N = 10^4, 10^5$ and $10^6$. B) Load as a function of the selection coefficient $s$, with $h = 0.01$ and $N = 10^6$.

**Supplementary figure A1.16:**

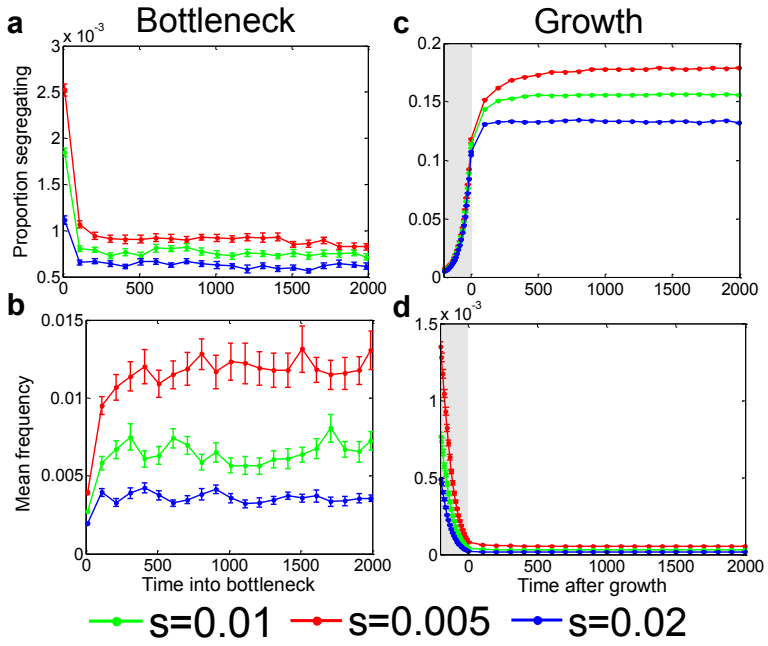
**Equilibrium properties of segregating sites in the quasi-dominant case**




Supplementary Fig. A1.16: The equilibrium properties of segregating sites in the quasi-dominant case. In all graphs, $h = 0.5$ and $u = 10^{-8}$. A) Frequency of deleterious alleles as a function of time in simulations with two population sizes, corresponding to $N = 10^4$ and $2 \cdot 10^4$. In both cases, $s = 0.01$. B) The expected proportion of segregating sites as a function of population size. C) The expected frequency of deleterious alleles at segregating sites as a function of population size.

**Supplementary figure A1.17:**

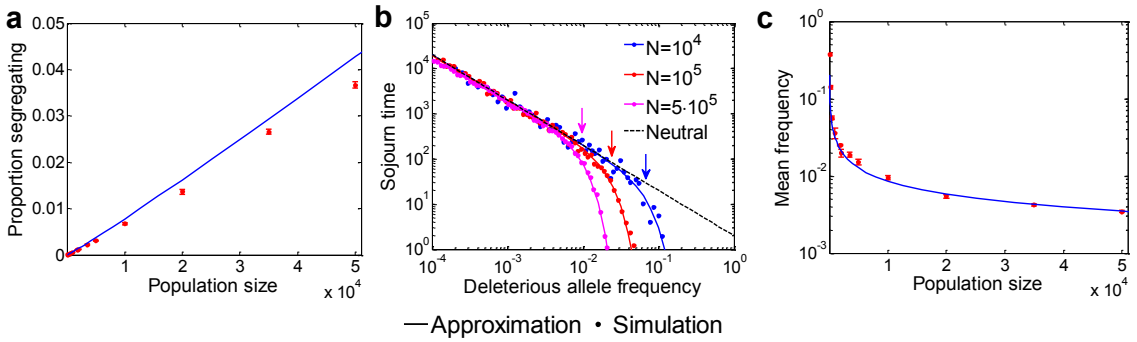
**Properties of segregating sites as a function of time for the quasi-dominant case**




Supplementary Fig. A1.17: The properties of segregating sites as a function of time for the quasi-dominant case. In all graphs, $h = 0.5$. The proportion of segregating sites after (A) the reduction in population size in the bottleneck model and (C) the onset of growth. The expected frequency of deleterious alleles at segregating sites after (B) the reduction in population size in the bottleneck model and (D) after the onset of growth. The shaded region is the period of growth in the Tennessen model.

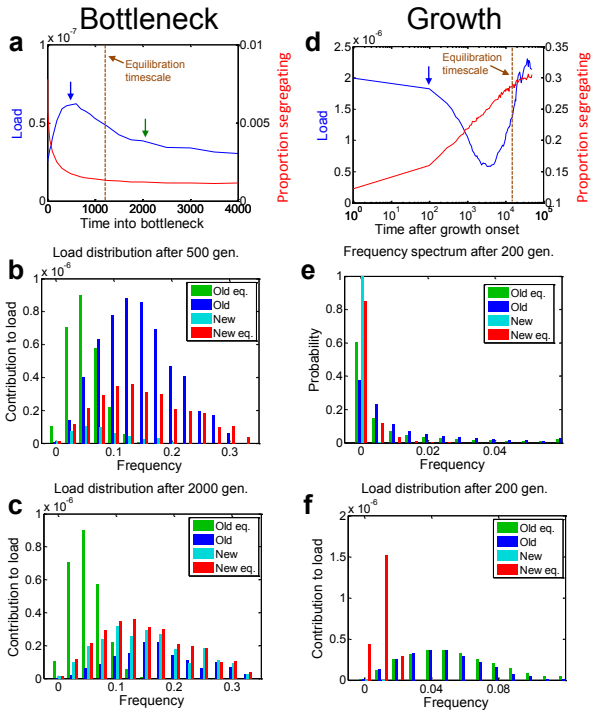
**Supplementary figure A1.18:**

**Properties of segregating sites at equilibrium in the recessive case**




Supplementary Fig. A1.18: The properties of segregating sites at equilibrium in the recessive case, as a function of population size. The selection coefficient is $s = 0.01$. (A) The proportion of segregating sites.(B) The sojourn time of deleterious alleles for different population sizes. The threshold frequency of $\frac{1}{\sqrt{2Ns}}$ for each population size is marked by an arrow with the corresponding color. (C) The average frequency of deleterious alleles.

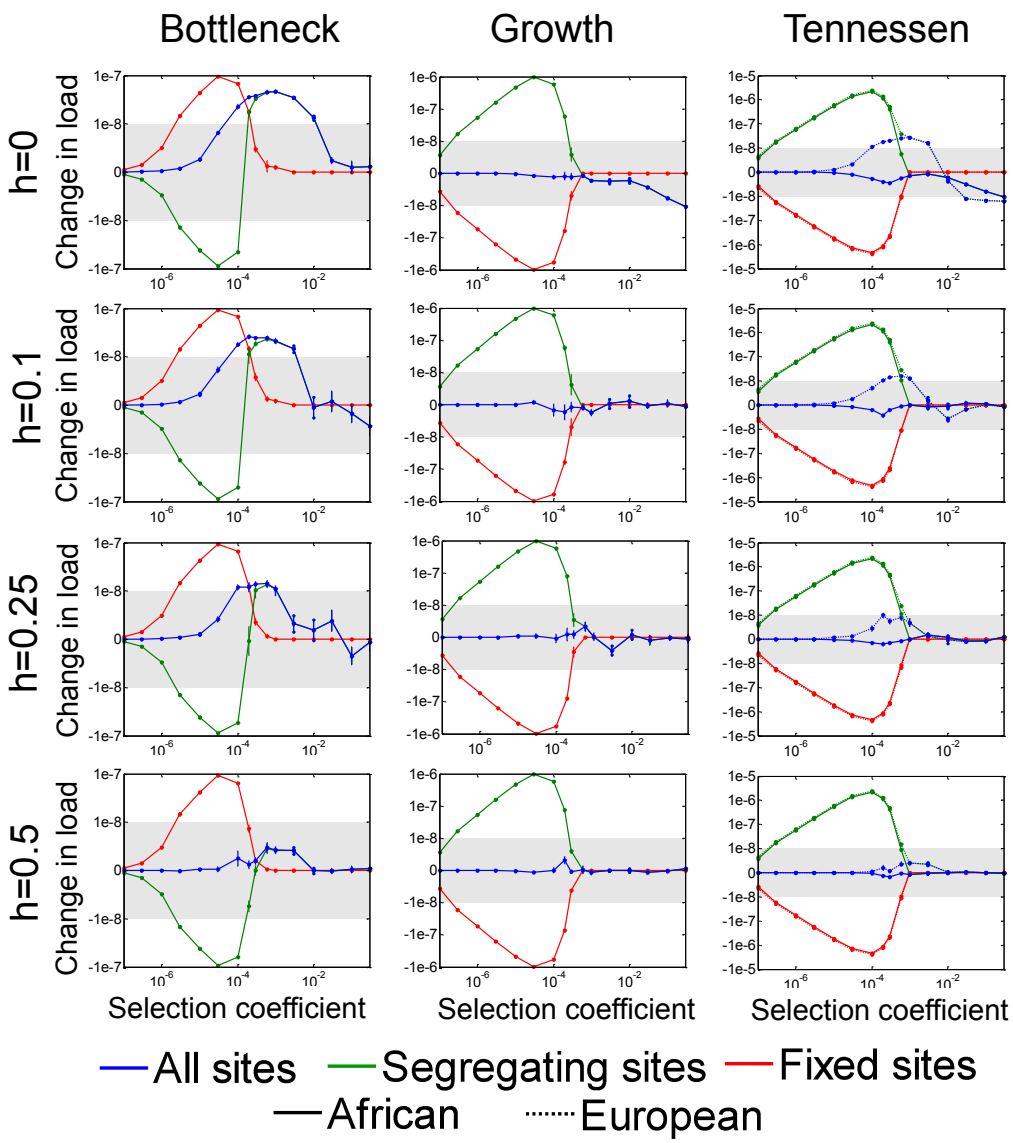
**Supplementary figure A1.19:**
**Load as a function of time in the recessive case**



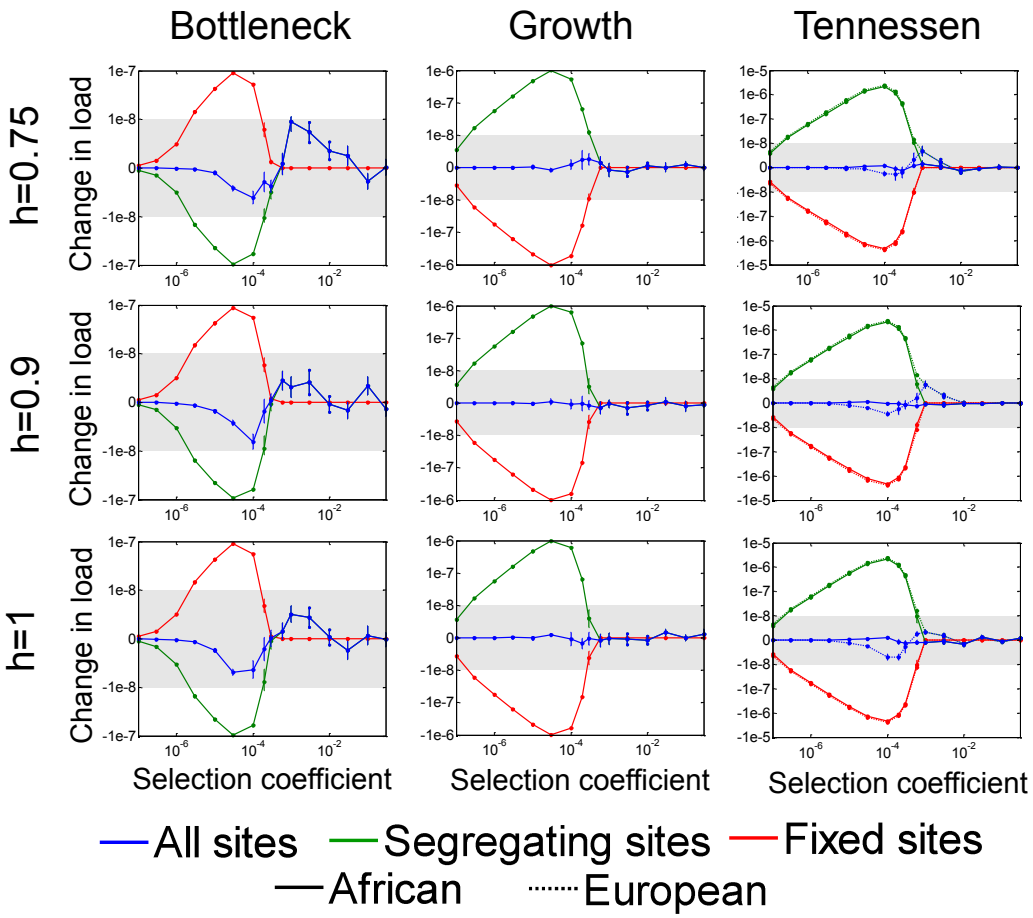

Supplementary Fig. A1.19: Load as a function of time in the recessive case. The selection coefficient is $s = 0.01$. A) The load and proportion of segregating sites as a function of time after the reduction in population size. B) The contribution to load of old and new mutations as a function of frequency, at the time of peak load (500 generations after the reduction in population size, indicated by a blue arrow in A). C) Same as B but for the time since the Out-of-Africa bottleneck, i.e., 50Kya (indicated by a green arrow in A). D) The load and proportion of segregating sites as a function of time after the onset of growth. E) The allele frequency distribution of old and new mutations at the end of the growth period (200 generations after onset, indicated by an arrow in D). F) The contribution to load of old and new mutations as a function of frequency at the end of the growth period.

**Supplementary figure A1.20:**

**Changes in load under the three demographic models with different dominance coefficients**




Supplementary Fig. A1.20: Continued on the next page.




Supplementary Fig. A1.20: Changes in load under the three demographic models with different dominance coefficients. $h = 0$ and $1/2$ correspond to the results in Supplementary Figure A1.10 and are provided for comparison.

# References


1. Charlesworth, B. & Charlesworth D. *Elements of Evolutionary Genetics* (Roberts and Co., 2010).
2. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69 (2012).
3. Ewens, W. J. *Mathematical Population Genetics*, 2nd ed. (Springer, 2004).
4. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-220 (2013).
5. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
6. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nature Genetics* **44**, 1161–1165 (2012).
7. Gutenkunst, R. N., Hernandez R. D., Williamson S. H. & Bustamante C. D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet.* **5**, e1000695 (2009).
8. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
9. Tajima, F. The effect of change in population size on DNA polymorphism. *Genetics* **123**, 597–601 (1989).
10. Hudson, R. R. Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44 (1990).
11. Wakeley, K. *Coalescent Theory: An Introduction* (Roberts and Co., 2008).
12. Gillespie, J. H. *Population Genetics: A Concise Guide*, 2nd ed. (Johns Hopkins University Press, 2004).
13. Feller, W. *An Introduction to Probability Theory and its Applications* (Wiley, 1968).
14. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710-722 (2010).
15. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740-743 (2012).

16. Otto, S. P. & Whitlock, M. C. The probability of fixation in populations of changing size. *Genetics* **146**, 723-733 (1997).

17. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **1**, 131 (2010).

18. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100-104 (2012).

19. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124-137 (2001).

20. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl. Acad. Sci. USA* **107**, 1752-1756 (2010).

21. Wang, K., Li, M. &Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

22. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248-249 (2010).

23. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4**, 1073-1081 (2009).

24. Schwarz, J.M., Rödelsperger, C., Schuelke, M. &Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575-576 (2010).

25. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553-1561 (2009).

# Appendix 2

## Contents

# 1. The model

## 1.1. Absorbing the environmental contribution into the fitness function

Here, we show that the additive environmental contribution to the phenotype can be absorbed into the fitness function, which justifies considering only the additive genetic contribution in our analysis. This result has been derived multiple times for the one dimensional case (e.g., 1). The argument in the multi-dimensional case is similar and included for completeness.

First, assume that the additive environmental contribution to the phenotype, $\vec{r}_e$, is distributed as a multi-normal with mean 0 and isotropic variance $V_e$. The expected absolute fitness of an individual with additive genetic contribution to the phenotype, $\vec{r}_g$, is given by averaging fitness over the distribution of environmental contributions. Namely,

$$\overline{\mathrm{W}}(\vec{r}_g) = \int_{\vec{r}_e} \frac{1}{(2\pi V_e)^{n/2}}\, e^{-\frac{\|\vec{r}_e\|^2}{2V_e}}\, \mathrm{W}(\vec{r}_g + \vec{r}_e) = \int_{\vec{r}_e} \frac{1}{(2\pi V_e)^{n/2}}\, e^{-\frac{\|\vec{r}_e\|^2}{2V_e}}\, e^{-\frac{\|\vec{r}_g+\vec{r}_e\|^2}{2w^2}}$$

$$= \frac{1}{(1+V_e/w^2)^{n/2}}\, e^{-\frac{\|\vec{r}_g\|^2}{2(w^2+V_e)}}. \qquad \text{(A2.1)}$$

Given that absolute fitness is defined up to a multiplicative constant, we can therefore absorb the additive environmental contribution by using the Gaussian fitness function

$$\widetilde{W}(\vec{r}_g) = \exp\left(-\frac{\|\vec{r}_g\|^2}{2\widetilde{w}^2}\right), \quad \text{(A2.2)}$$

where $\widetilde{w}^2 = w^2 + V_e$. Even when the environmental contribution is anisotropic, we can always choose a coordinate system in which the effective fitness function takes an isotropic form around the fitness peak (Eq. 1, which appears in the Model section of Chapter 2).

### 1.2. The distribution of mutational effect sizes on a given trait

In Chapter 2, we define the distribution of phenotypic effects of newly arising mutations in the *n*-dimensional trait space, $\vec{a}$. Here, we consider the projection of these effects on a given trait, $a_1$, taken without loss of generality to be on the 1st dimension. The distribution of effect sizes on a focal trait will depend on the degree of pleiotropy, *n*, and the form of this dependency becomes important when we consider how pleiotropy affects genetic architecture.

We want to calculate the distribution of effect sizes on the focal trait, $a_1$, conditional on their overall effect, $a = \|\vec{a}\|$. We assume that the distribution of effects of de novo mutations is isotropic in trait space. The effect of a mutation, $\vec{a}$, therefore has equal probability to occupy any point on an *n*-dimensional sphere with radius $a$. Let $S_m(x)$ denote the surface area of an *m*-dimensional sphere of radius $x$ and $\theta$ denote the angle between the vector $\vec{a}$ and its projection $a_1$, i.e., $a_1 = a\cos\theta$. In these terms, the surface area element corresponding to angle $\theta$ is

$$S_{n-1}(a\sin\theta)\, a\, d\theta, \quad \text{(A2.3)}$$

and by a change of variables, the surface area element corresponding to projection $a_1$ on the focal trait is

$$\mathrm{S}_{n-1}(a\sin\theta)\,a\,d\theta = \mathrm{S}_{n-1}\left(\sqrt{a^2-a_1^2}\right)\frac{a}{\sqrt{a^2-a_1^2}}da_1, \tag{A2.4}$$

since $da_1 = \left|\frac{da_1}{d\theta}\right| d\theta = a\sin\theta\, d\theta = \sqrt{a^2-a_1^2}d\theta$. This result implies that the probability density of $a_1$ is

$$\varphi_n(a_1|a) = \frac{\mathrm{S}_{n-1}\left(\sqrt{a^2-a_1^2}\right)}{\mathrm{S}_n(a)}\frac{a}{\sqrt{a^2-a_1^2}} = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\left(1-\frac{a_1^2}{a^2}\right)^{\frac{n-3}{2}}\frac{1}{a} \tag{A2.5}$$

(for a similar derivation, see (2)).

Next, we consider the high pleiotropy limit form of this distribution. For any degree of pleiotropy, the symmetry of the mutational distribution implies that

$$\mathrm{E}(a_1|a) = 0 \tag{A2.6}$$

and the equivalence among traits implies that

$$\mathrm{V}(a_1|a) = a^2/n \tag{A2.7}$$

(see Chapter 2 for more details). It follows that when *n* becomes sufficiently large, $a_1/a \ll 1$ and therefore

$$\left(1-\frac{a_1^2}{a^2}\right)^{\frac{n-3}{2}} \approx \exp\left(-\frac{n}{2}\,\frac{a_1^2}{a^2}\right). \tag{A2.8}$$

In addition, $\Gamma\left(\frac{n}{2}\right)/\Gamma\left(\frac{n-1}{2}\right) \approx \sqrt{n/2}$. Substituting these expressions into Eq. A2.5, we find that for sufficiently large *n* the distribution of effect sizes approaches the normal distribution

$$\varphi_n(a_1|a) \approx \frac{1}{\sqrt{2\pi(a^2/n)}}\exp\left(-\frac{1}{2}\,\frac{a_1^2}{a^2/n}\right). \tag{A2.9}$$

As we elaborate in Chapter 2, important implications about quantitative genetic variation follow from this high pleiotropy limit. The limit also holds quite generally when the distribution of effect sizes is anisotropic (see Section 5.4).

## 2. Solving for summaries of genetic architecture

Here, we derive closed forms for summaries of genetic architecture under our model. We begin by deriving the first two moments of change in allele frequency in a single generation. With these moments at hand, we use the diffusion approximation to calculate the sojourn time for alleles that contribute to quantitative genetic variation (3). Together with the distribution of effect sizes derived in the previous section, the sojourn time allows us to obtain closed forms for summaries of genetic architecture. Specifically, we can obtain a closed form for any summary that can be described as a function of allele frequencies and effect sizes at sites contributing to quantitative genetic variation. We use these expressions to calculate the summaries used in Chapter 2, for example the expected additive genetic variance and its distribution across sites.

### 2.1. The first two moments of change in allele frequency

We assume that:

- The phenotypic distribution at steady state is well approximated by an isotropic multivariate normal distribution centered at the optimum, namely by the probability density

  $$f(\vec{r}) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left(-\frac{r^2}{2\sigma^2}\right). \quad \text{(A2.10)}$$

- Both $a^2$ and $\sigma^2 \ll w^2$.

These assumptions are justified in Section 3.3.

We rely on these assumptions to calculate the first two moments of change in frequency in a single generation for an allele with phenotypic effect $\vec{a}$ and frequency $q$. The fitnesses of the three genotypes at the site depend on its distribution of genetic backgrounds, i.e., on the total phenotypic contribution of sites other than the focal one, $\vec{R}$. Following Eq. A2.10 and assuming

every allele contributes only a small proportion of the genetic variance, the distribution of $\vec{R}$ is well approximated by

$$\mathrm{f}\left(\vec{R}\middle|\vec{a},q\right)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\left\|\vec{R}+2q\vec{a}\right\|^2}{2\sigma^2}\right). \quad \text{(A2.11)}$$

The expected fitnesses of the three genotypes then follow from integrating over backgrounds:

$$W_{00}=\int_{\vec{R}}\mathrm{f}\left(\vec{R}\middle|\vec{a},q\right)\mathrm{W}\left(\vec{R}\right)=\left(\frac{w}{\sqrt{w^2+\sigma^2}}\right)^n\exp\left(-\frac{4a^2q^2}{2(w^2+\sigma^2)}\right), \quad \text{(A2.12)}$$

$$W_{01}=\int_{\vec{R}}\mathrm{f}\left(\vec{R}\middle|\vec{a},q\right)\mathrm{W}\left(\vec{R}+\vec{a}\right)=\left(\frac{w}{\sqrt{w^2+\sigma^2}}\right)^n\exp\left(-\frac{4a^2(q-1/2)^2}{2(w^2+\sigma^2)}\right) \quad \text{(A2.13)}$$

and

$$W_{11}=\int_{\vec{R}}\mathrm{f}\left(\vec{R}\middle|\vec{a},q\right)\mathrm{W}\left(\vec{R}+2\vec{a}\right)=\left(\frac{w}{\sqrt{w^2+\sigma^2}}\right)^n\exp\left(-\frac{4a^2(q-1)^2}{2(w^2+\sigma^2)}\right). \quad \text{(A2.14)}$$

The first moment of change in allele frequency is then

$$\mathrm{E}(\Delta q)=-pq\frac{p(W_{00}-W_{01})+\mathrm{q}(W_{01}-W_{11})}{\bar{W}}\approx-\frac{a^2}{w^2}pq\left(\frac{1}{2}-q\right), \quad \text{(A2.15)}$$

relying on our assumptions that $a^2$ and $\sigma^2 \ll w^2$. The functional form of the first moment is equivalent to that of the standard viability selection model with under-dominance and selection coefficient $s=\frac{a^2}{w^2}$ or scaled selection coefficient

$$S=2N\frac{a^2}{w^2}. \quad \text{(A2.16)}$$

Similarly, we find that

$$\mathrm{V}(\Delta q)\approx\frac{pq}{2N}, \quad \text{(A2.17)}$$

which is the standard second moment with genetic drift.

## 2.2. Sojourn time

Based on the first two moments, we can use the diffusion approximation to calculate the sojourn time as a function of allele frequency, i.e., the density of the time that an allele spends at a given

frequency $q$ before it fixes or is lost (3). For a mutant allele with initial frequency $1/2N$ and scaled selection coefficient *S*, the sojourn time is

$$\tau(q|S) = \begin{cases} \frac{N\sqrt{\pi/S}}{\mathrm{erf}(\sqrt{S}/2)} \; \frac{e^{S(1-2x)^2/4}}{x(1-x)} \, \mathrm{f}_-(S,q)\mathrm{f}_+(S,1/2N) & 0 \le q \le 1/2N \\ \frac{N\sqrt{\pi/S}}{\mathrm{erf}(\sqrt{S}/2)} \; \frac{e^{S(1-2x)^2/4}}{x(1-x)} \, \mathrm{f}_+(S,q)\mathrm{f}_-(S,1/2N) & 1/2N \le q \le 1 \end{cases} \quad \text{(A2.18)}$$

where erf is the error function and $\mathrm{f}_{\pm}(S,y) \equiv \mathrm{erf}\left(\sqrt{S}/2\right) \pm \mathrm{erf}\left(\sqrt{S}(1-2y)/2\right)$.

The sojourn time takes simple limiting forms when selection is effectively neutral ($\square \ll 1$) or strong ($S \gg 1$). In the effectively neutral range, it is well approximated by $\tau(q|S) = 2/\mathrm{q}$, and in the strongly selected range, it is well approximated by $\tau(q|S) = 2\exp(-Sq)/q$.

### 2.3. Calculating expectations of summaries of architecture

Many summaries of interest can be expressed as sums over segregating sites of some function $c(q, a_1)$, where $q$ is the derived allele frequency and $a_1$ is the effect size on the trait. For example, the additive genetic variance in a trait is given by the sum of $v(q, a_1) = 2a_1^2 q(1-q)$ over sites. The expectation over such summaries can be expressed as

$$\mathrm{E}(C) = 2NU \int_q \int_{a_1} c(q, a_1)\rho(q, a_1) \, , \quad \text{(A2.19)}$$

where $C$ is the summery summed over all sites, $2NU$ is the population mutation rate per generation and $\rho(q, a_1)$ is the density of sites with the corresponding frequency and effect size per unit mutational input.

The density $\rho(q, a_1)$ can be broken down into contributions from sites with different selection coefficients, i.e.,

$$\rho(q, a_1) = \int_S \mathrm{f}(S)\big(\tau(q|S)\eta(a_1|S)\big), \quad \text{(A2.20)}$$

where $\mathrm{f}(S)$ is the distribution of selection coefficients and $\tau(q|S)$ is the sojourn time of a mutation with selection coefficient $S$ (Eq. A2.18). The probability density $\eta(a_1|S)$ of effect sizes given selection coefficient $S$ follows from Eqs. A2.5 and A2.16

$$\eta(a_1|S) = \varphi_{\mathrm{n}}(a_1|a(S)) = \varphi_{\mathrm{n}}\left(a_1\middle|\sqrt{(w^2/2N)S}\right). \quad \text{(A2.21)}$$

This allows us to break down our summaries into contributions from sites with different selection coefficients

$$\mathrm{E}(C) = 2NU\int_S \mathrm{f}(S)\mathrm{E}(C|S) \quad \text{(A2.22)}$$

with

$$\mathrm{E}(C|S) = \int_q \int_{a_1} c(q,a_1)\tau(q|S)\eta(a_1|S). \quad \text{(A2.23)}$$

We use Eq. A2.23 to study how summaries of architecture depend on the strengths of selection, and how these summaries will depend on different distributions of selection coefficients. This allows us to draw general implications about genetic architecture despite our limited knowledge about this distribution.

## 3. Additive genetic variance and number of segregating sites

The distributions of additive genetic variance and of the number of segregating sites are critical to understanding genetic architecture and specifically to interpreting results of GWAS. Here we derive closed forms for both distributions as well as simple approximations under strong and effectively neutral selection.

### 3.1. Expectations

We begin by considering the expected contribution of a site to additive genetic variance. Substituting the contribution to variance from a single site $v(q, a_1) = 2a_1^2 q(1-q)$ into Eq. A2.23, we find that

$$\mathrm{E}(V|S) = \int_q \int_{a_1} v(q, a_1)\tau(q|S)\eta(a_1|S) = \int_q 2q(1-q)\tau(q|S) \int_{a_1} a_1^2 \eta(a_1|S)$$

$$= \frac{2w^2}{nN} \int_q \frac{1}{2} Sq(1-q)\tau(q|S)\,. \qquad \text{(A2.24)}$$

The total additive genetic variance over all sites is

$$\sigma^2 = 2NU \int_S \mathrm{f}(S)\mathrm{E}(V|S). \qquad \text{(A2.25)}$$

The closed form for $\mathrm{E}(V|S)$ in Eq. A2.24 was integrated numerically to obtain Fig. 2a in Chapter 2. We can use the results of Keightley and Hill (4) to obtain an analytic approximation for $\mathrm{E}(V|S)$:

$$\mathrm{E}(V|S) = \frac{2w^2}{nN} \sqrt{\frac{\pi S}{4}} \mathrm{erfi}\left(\sqrt{S}/4\right) \exp(-\mathrm{S}/4) + \mathrm{O}\left(\frac{1}{2N}\right), \qquad \text{(A2.26)}$$

where erfi is the imaginary error function ($\mathrm{erfi}(x) \equiv \mathrm{erf}(ix)\,/i$).

In the effectively neutral and strong selection limits, we can use limit forms of the sojourn time to derive simple approximations for $E(V|S)$. In the effectively neutral limit, i.e., when $S \ll 1$, $\tau(q|S) \approx 2/q$ and therefore

$$\mathrm{E}(V|S) \approx \frac{2w^2}{nN}\frac{S}{2}. \qquad \text{(A2.27)}$$

In practice, this expression provides a decent approximation when $S < 1$ (Fig. A2.1a). In the strong selection limit, when $S \gg 1$, $\tau(q|S) \approx 2\exp(-Sq)/q$ and therefore

$$\mathrm{E}(V|S) \approx \frac{2w^2}{nN}. \qquad \text{(A2.28)}$$

In practice, this expression provides a decent approximation when $S > 30$ (Fig. A2.1a). The constant $2w^2/nN$, which recurs in our derivations (e.g., Eq. A2.24), thus has a simple interpretation: it is the expected contribution of strongly selected sites to additive genetic variance (per unit mutational input). We therefore denote it by $v_s$, and henceforth measure variance in these units.

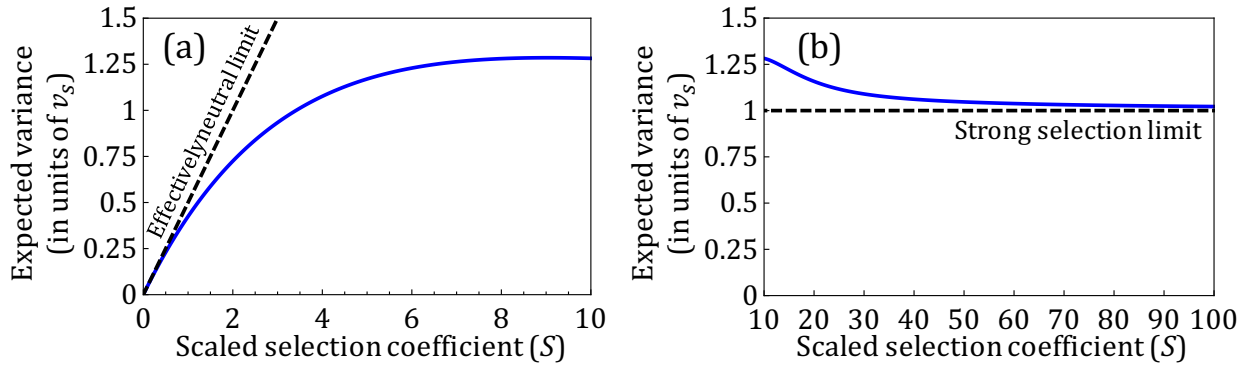



**Figure A2.1.** The effectively neutral and strong selection approximations for the expected contribution to genetic variance per site. (A) The expression in the limit of $S \ll 1$ provides a decent approximation when $S < 1$. (B) The expression in the limit of $S \gg 1$ provides a decent approximation when $S > 30$.

We next consider how the expected number of segregating sites depends on the strength of selection. This expectation (per unit mutational input) is simply the mean sojourn time of a newly arising mutation. Formally, it follows from substituting $\mathrm{k}(q, a_1) = 1$ into Eq. A2.23, i.e.,

$$\mathrm{E}(K|S) = \int_q \int_{a_1} \tau(q|S)\eta(a_1|S) = \int_q \tau(q|S) \int_{a_1} \eta(a_1|S) = \int_q \tau(q|S). \qquad \text{(A2.29)}$$

In Fig. A2.2, we calculate this integral numerically for different values of $S$, to find that the number of segregating sites depends only weakly on $S$. Intuitively, this follows from the fact that the vast majority of mutations, be they effectively neutral, intermediate, or strongly selected, spend only a few generations in the population at low copy numbers before going extinct.

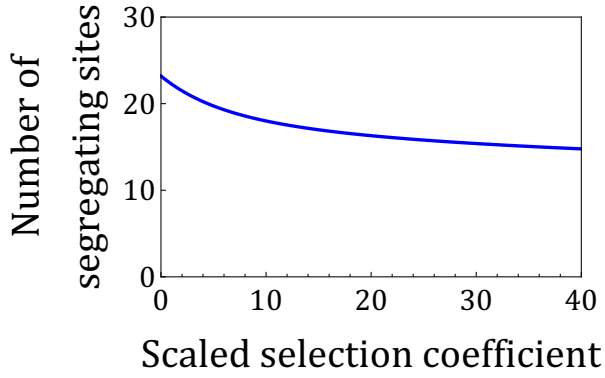



**Figure A2.2.** The number of segregating sites per unit mutational input (or, equivalently, the expected sojourn time of a newly arising mutation), Eq. A2.29, is only weakly dependent on the strength of selection. Calculated for a population size of 20,000.

### 3.2. Densities

Here, we consider how the additive genetic variance is distributed among sites. We begin by deriving a closed form for the density of segregating sites with a given contribution to variance $v$. This density follows from substituting Dirac's delta function $\delta\left(v - 2a_1^2 q(1-q)\right)$ into Eq. A2.23:

$$\rho(v|S) = \mathrm{E}\left(\delta\left(v - 2a_1^2 q(1-q)\right)\middle|S\right) = \int_q \int_{a_1} \delta\left(v - 2a_1^2 q(1-q)\right)\tau(q|S)\eta(a_1|S) =$$

$$= \int_{a_1} \left(\tau(\mathrm{q}_+(v,a_1)|S)\left|\frac{\partial \mathrm{q}_+(v,a_1)}{\partial v}\right| + \tau(\mathrm{q}_-(v,a_1)|S)\left|\frac{\partial \mathrm{q}_-(v,a_1)}{\partial v}\right|\right)\eta(a_1|S)$$

$$= \int_{a_1}\left(\tau(\mathrm{q}_+(v,a_1)|S) + \tau(\mathrm{q}_-(v,a_1)|S)\right)\frac{1}{2a_1^2\sqrt{1-2v/a_1^2}}\,\eta(a_1|S), \qquad \text{(A2.30)}$$

where $\mathrm{q}_\pm(v,a_1) = \frac{1}{2}\left(1 \pm \sqrt{1-2v/a_1^2}\right)$ are the two frequencies for which $v = 2a_1^2 q(1-q)$. This integral can be calculated numerically for any $S$ and degree of pleiotropy $n$ (by using the corresponding density $\eta(a_1|S)$). Moreover, as we illustrate below, summary statistics of the

distribution of variances among sites can be expressed and calculated in terms of integrals over the density $\rho(v|S)$.

We can greatly simplify the expression for $\rho(v|S)$ in the limits of effectively neutral and strong selection, and especially in the cases without pleiotropy or with extensive pleiotropy. When selection is effectively neutral ($S \ll 1$), then $\tau(q|S) \cong 2/q$ and thus

$$\rho(v|S) = \int_{a_1} \left(\frac{2}{\mathrm{q}_+(v,a_1)} + \frac{2}{\mathrm{q}_-(v,a_1)}\right) \frac{1}{2a_1^2\sqrt{1-2v/a_1^2}}\, \eta(a_1|S)$$

$$= \int_{a_1} \left(\frac{4}{1+\sqrt{1-2v/a_1^2}} + \frac{4}{1-\sqrt{1-2v/a_1^2}}\right) \frac{1}{2a_1^2\sqrt{1-2v/a_1^2}}\, \eta(a_1|S)$$

$$= \int_{a_1} \frac{2}{v\sqrt{1-2v/a_1^2}}\, \eta(a_1|\mathrm{S}), \qquad \text{(A2.31)}$$

with variance measured in units of $v_s$ and effect size measured in units of $\sqrt{v_s}$. Without pleiotropy ($n = 1$), the effect size is $a_1 = \pm\frac{1}{2}\sqrt{S}$ and the expression for the density simplifies to

$$\rho(v|S) = \frac{2}{v\sqrt{1-v/v_{max}}}, \qquad \text{(A2.32)}$$

where $v_{max} \equiv S/8$ is the maximal contribution to variance for a mutation with selection coefficient $S$, which is obtained when both alleles have frequency ½. When the degree of pleiotropy is high ($n \gg 1$), $a_1$ is approximately normally distributed with mean 0 and variance $S/4$ (Eq. 11, which appears in the Results section of Chapter 2) and the expression for the density simplifies to

$$\rho(v|S) = \int_{a_1>\sqrt{2v}} \frac{2}{v\sqrt{1-2v/a_1^2}} \frac{2}{\sqrt{2\pi S/4}} \exp(-2a_1^2/S) = 2\exp(-4v/S)/v. \qquad \text{(A2.33)}$$

When selection is strong, derived alleles are rare ($q \ll 1$), implying that $v \ll a_1^2$ and $q \approx v/2a_1^2$, and that the sojourn time is well approximated by $\tau(q|S) = 2\exp(-Sq)/q$. The density $\rho(v|S)$ then simplifies to

$$\rho(v|S) \approx \int_{a_1} \tau(q(v,a_1)|S)\frac{1}{2a_1^2}\,\eta(a_1|S) \approx \int_{a_1} \frac{4a_1^2}{v}\exp(-Sv/2a_1^2)\frac{1}{2a_1^2}\,\eta(a_1|S)$$

$$= \int_{a_1} \frac{2}{v}\exp(-Sv/2a_1^2)\,\eta(a_1|S). \quad \text{(A2.34)}$$

Without pleiotropy, this expression further simplifies to

$$\rho(v|S) \approx 2\exp(-2v)/v, \quad \text{(A2.35)}$$

and when the degree of pleiotropy is high ($n \gg 1$), then

$$\rho(v|S) \approx \int_{a_1} \frac{2}{v}\exp(-Sv/2a_1^2)\frac{2}{\sqrt{2\pi S/4}}\exp(-2a_1^2/S) = 2\exp\left(-2\sqrt{v}\right)/v. \quad \text{(A2.36)}$$

We are especially interested in the distribution of variances among sites that exceed some threshold contribution $v^*$. As we discuss in Chapter 2 and in Section 6, to a first approximation, the loci identified in a GWAS would be those with contributions to additive variance that exceed the study's threshold contribution $v^*$. In particular, our inferences based on GWAS data rely on fitting the probability density of the number of segregating sites with variance $v$ that exceed a given threshold contribution $v^*$ (Section 7). This probability density is:

$$\mathrm{f}(v|S) = \rho(v|S)/\mathrm{K}(v^*|S), \quad \text{(A2.37)}$$

where

$$\mathrm{K}(v^*|S) \equiv \int_{v>v^*} \rho(v|S) \quad \text{(A2.38)}$$

is the expected number of segregating sites with contributions to variance exceeding $v^*$ per unit mutational input.

In our analysis, we focus on the expected proportion of additive genetic variance arising from sites that exceed a threshold contribution $v^*$, which approximates the heritable variance explained in GWAS. This proportion is given by

$$\mathrm{G}(v^*|S) = \frac{\int_{v>v^*} v\,\rho(v|S)}{\int_v v\,\rho(v|S)} = \frac{\int_{v>v^*} v\,\rho(v|S)}{E(V|S)}. \quad \text{(A2.39)}$$

Given a distribution of selection coefficients, $\mathrm{f}(S)$, the corresponding proportion is

$$\mathrm{G}(v^*) = \frac{\int_S \mathrm{G}(v^*|S)\mathrm{E}(V|S)\mathrm{f}(S)}{\int_S \mathrm{E}(V|S)\mathrm{f}(S)} = \int_S \mathrm{G}(v^*|S)\frac{\mathrm{E}(V|S)\mathrm{f}(S)}{\int_S \mathrm{E}(V|S)\mathrm{f}(S)}. \quad \text{(A2.40)}$$

The dependences of the proportion of variance $\mathrm{G}(v^*|S)$ and the number of sites $\mathrm{K}(v^*|S)$ on the strength of selection $S$ for cases without pleiotropy and with extensive pleiotropy are shown in Figs. 3 and A2.3, respectively. We rely on Eqs. A2.32, A2.33, A2.35, A2.36, A2.38 and A2.39 to derive simplified forms for these summaries in the effectively neutral and strongly selected limits (Table A2.1). While the expressions for the effectively neutral limit were derived for $S \ll 1$, in practice they provide a decent approximation when $S < 1$ (Fig. A2.4a & b). In the strong selection limit ($S \gg 1$), the expressions for the case without pleiotropy provide a decent approximation for $S > 30$ (Fig. 3a), whereas with extensive pleiotropy they already work quite well for $S > 5$ (Fig. 3b).

| Selection | Effectively neutral ($S \ll 1$) | | Strongly selected ($S \gg 1$) | |
|---|---|---|---|---|
| # of traits | $n = 1$ | $n \gg 1$ | $n = 1$ | $n \gg 1$ |
| $\mathrm{E}(V\|S)$ | $S/2$ | | 1 | |
| $\mathrm{G}(v^*\|S)$ | $\sqrt{1-8v^*/S}$ | $\exp(-4v^*/S)$ | $\exp(-2v^*)$ | $\left(1+2\sqrt{v^*}\right)\exp(-2\sqrt{v^*})$ |
| $\mathrm{K}(v^*\|S)$ | $4\cdot \mathrm{artanh}(\sqrt{1-8v^*/S})$ | $2\cdot \mathrm{I}(4v^*/S)$ | $2\cdot \mathrm{I}(2v^*)$ | $4\cdot \mathrm{I}(2\sqrt{v^*})$ |

**Table A2.1.** Limits for the expected proportion of variance and expected number of sites corresponding to sites that exceed a threshold contribution to additive genetic variance $v^*$. $\mathrm{I}(x) \equiv \int_{t>x} \exp(-t)/t$ is an exponential integral and artanh is the inverse hyperbolic tangent.

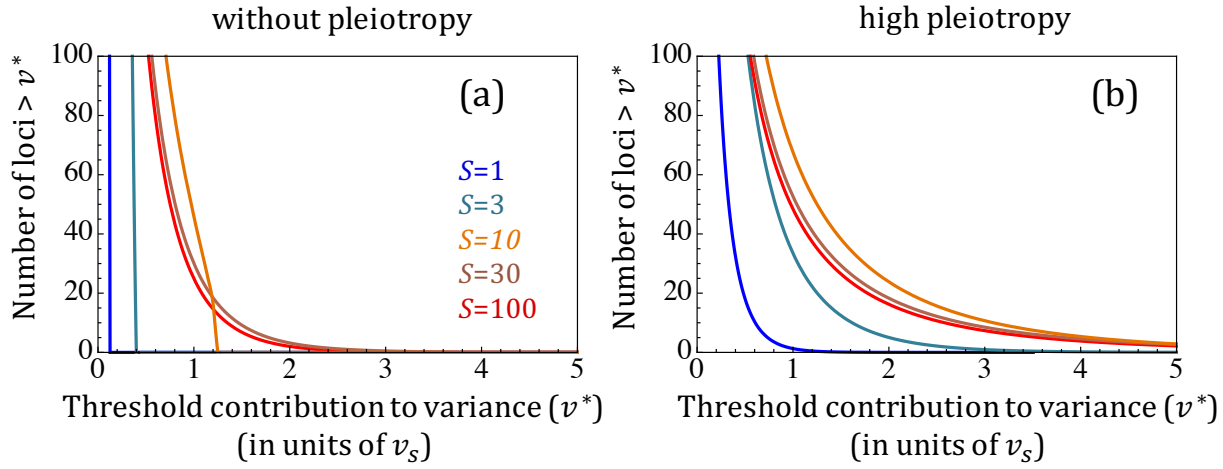



**Figure A2.3.** The number of loci per Mb contributing more than $v^*$ to the variance, as a function of $v^*$, in the case without pleiotropy, $n = 1$ (a), and in the high pleiotropy limit, $n \gg 1$ (b). We assume a constant population size of 20,000, with a mutation rate of $1.2 \cdot 10^{-8}$ (5).

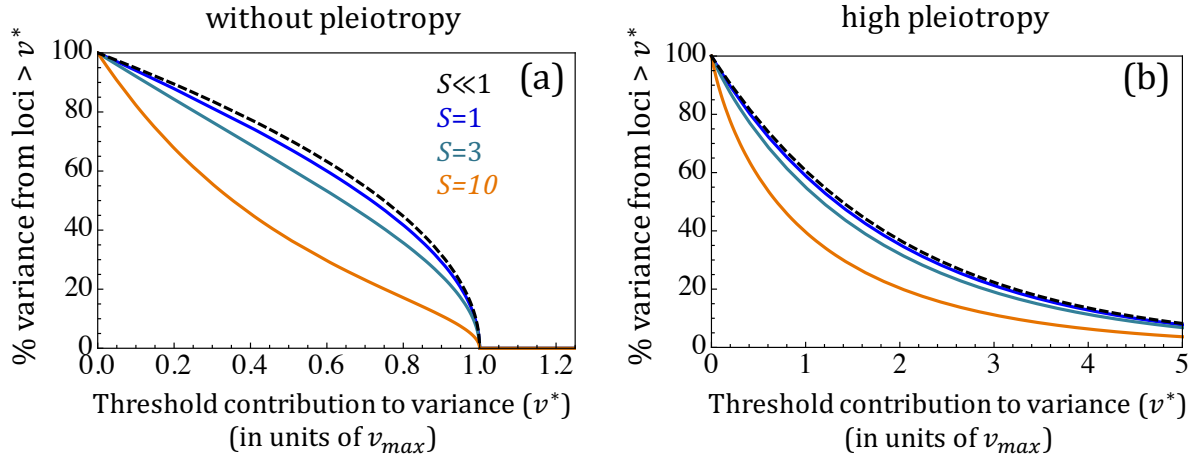



**Figure A2.4.** The proportion of additive genetic variance that arises from sites that contribute more than the value on the x-axis, for a single trait (a) and in the pleiotropic limit (b). We show the x-axis in units of $v_{max} = (S/8) \cdot v_s$ in order to evaluate the approximations in the effectively neutral limit (in dashed black; Eqs. 14 & 16, which appear in the Results section of Chapter 2); note that $v_{max}$ is not the maximal variance in cases with pleiotropy.

Both the proportion of variance, $G(v^*|S)$, and number of variants, $K(v^*|S)$, appear to always increase with the degree of pleiotropy, $n$ (Fig. A2.5). We do not have a proof for this property but can suggest an intuitive explanation. Without pleiotropy ($n$=1), the selection coefficient determines the effect size, such that any contribution $v^*$ to genetic variance corresponds to a specific minor allele frequency $q^*$. The sites with contributions $v > v^*$ are therefore those with minor allele frequencies $q > q^*$. Pleiotropy causes sites with a given selection coefficient to have a distribution of effect sizes on the trait under consideration. As a result, some sites with frequencies above $q^*$ end up with contributions to variance below $v^*$ while others exceed $v^*$. To understand how this affects $G(v^*|S)$, recall that for any selection coefficient, the density of variants always rapidly increases as $q^*$ decreases. As long as the contribution $v^*$ and the corresponding frequency without pleiotropy $q^*$ are not close to 0, we may therefore expect that introducing pleiotropy would result in pushing more sites above $v^*$ than below $v^*$, resulting in a net increase to the proportion $G(v^*|S)$. For the same reasons, the number of variants with $v > v^*$ shows a similar behavior and also grows with $n$.

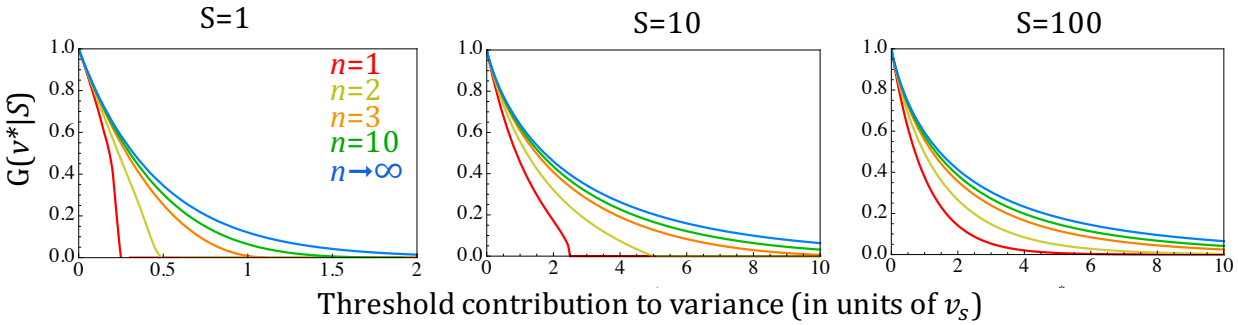



**Figure A2.5.** The effect of pleiotropy on the proportion of the variance explained by sites contributing more then $v^*$ to the variance, $G(v^*|S)$ (see Eq. A2.39). For all selection coefficients, the proportion of variance explained increases as the number of traits, i.e., the degree of pleiotropy, increases.

### 3.3. Comparing predictions against simulations

We tested our theoretical derivations for the total genetic variance and its distribution among sites against forward computer simulations. The code and documentation are available at https://github.com/sellalab/GenArchitecture. The simulation implements the model as specified in Chapter 2, with the following additional details and one exception. First, we assume the infinite sites model for mutation. Second, the distribution of scaled selection coefficients, or equivalently the distribution of mutation sizes (see Eq. 7, which appears in the Results section of Chapter 2), is taken to be a Gamma distribution, with specified parameters (see below). For computational efficiency, we use fecundity rather than viability selection; however, we ran a smaller number of simulations to verify that this choice does not lead to a detectable difference in the results. Each simulation is run for a burn-in period of $10N$ generations, to ensure convergence to the steady state behavior, before the variances at segregating sites are measured.

We explore a range of parameter values chosen to balance biological plausibility (see Section 4) and manageable running times. Notably, we used a population size of $N = 1000$, with a burn-in time of 10,000 generations. We vary the number of traits to include $n = 1, 3, \& 10$, and vary the mutation rate per haploid genome per generation within the range $1/2\mathrm{N} \leq U \leq 1$ (see Section 4), including $U$=0.0005, 0.001. 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5 & 1. Selection coefficients are chosen from an exponential distribution (setting the shape parameter for the Gamma distribution to 1) with means E($S$) = 0.1, 10, and 50. For simplicity, we take $w^2 = 1$, which is equivalent to choosing the units used to measure effect sizes.

The simulation results for the total genetic variance and its distribution among sites are in close agreement with our theoretical predictions (Fig. A2.6). Specifically, within the parameter ranges that we assume, i.e., when $1/2N \ll \sigma^2/w^2 \ll 1$ (see Section 4), the total genetic variances

measured in simulations are indistinguishable from our predictions (Fig. A2.6a). Moreover, simulations and prediction seem to agree even when $\sigma^2/w^2 \leq 1/2\mathrm{N}$, although we consider this range to be less relevant, given our focus on highly polygenic traits (see Section 4.4). We also compare simulated and predicted distributions of variances among sites, in terms of $\mathrm{G}(v)$, the proportion of the variance arising from sites that contribute more than $v$ (Eq. A2.40), and find them to be in close agreement (Fig. A2.6b).

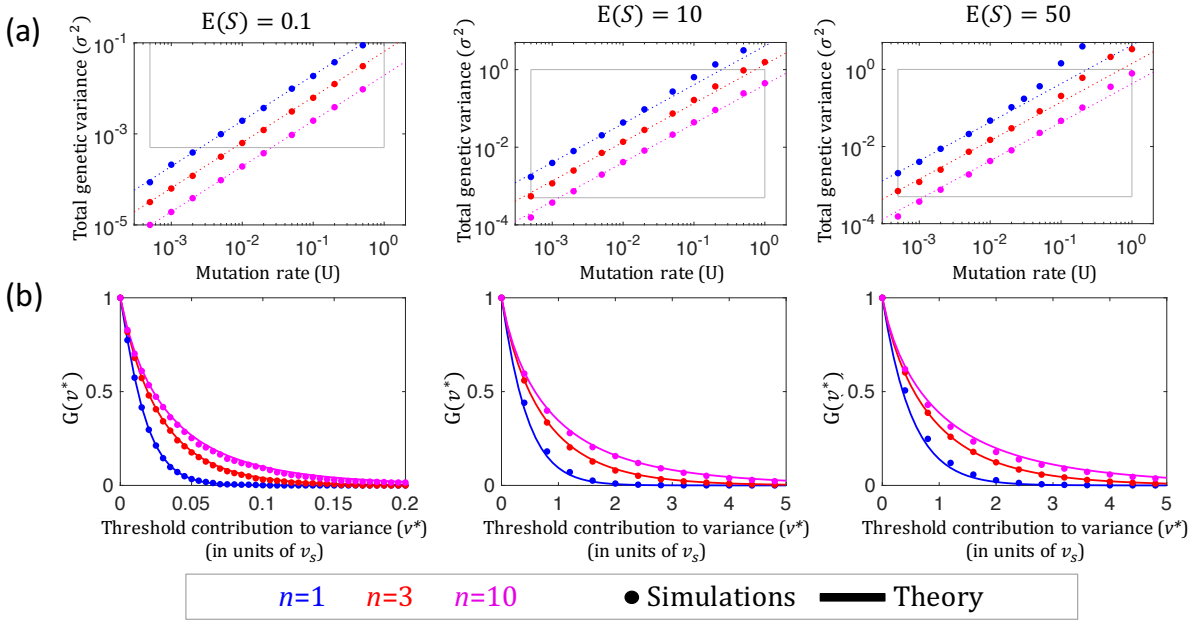



**Figure A2.6.** Testing theoretical predictions for genetic variance against simulation results. (a) Total genetic variance (in units of $w^2$) as a function of the mutation rate. The biologically relevant range of mutation rates, $1/2N \ll U \ll 1$, and the range in which we expect our predictions to be valid, $w^2/2N \ll \sigma^2 \ll w^2$, are marked by a grey box. (b) The distribution of variances among sites, for $U = 0.01$; $\mathrm{G}(v^*)$ is the proportion of variance from sites contributing more than $v^*$ (Eq. A2.40). Error bars represent one standard deviation. For each set of parameters, the number of simulations was chosen to obtain standard deviations below 10%. In practice, we often obtain much smaller standard deviations, which is why most error bars are too small to be visible.

We ran two additional variations on the basic simulation procedure (also available at https://github.com/sellalab/GenArchitecture): one to explore the effects of a shift in the optimal phenotype (Section 5.1 and Fig. A2.7) and another to explore the effects of asymmetric

mutational input (Section 5.2 and Fig. A2.8). To these ends, for simplicity, we consider the case without pleiotropy, i.e., with $n = 1$. In the first, after the $10N$ generations burn-in period, we introduce a shift in the optimal phenotype, and trace the allelic behavior over an additional 4,000 generations (see Section 5.1). In the second, after the $10N$ generations burn-in period, we introduce asymmetry in the rates of trait increasing and decreasing mutations, and trace the allelic trajectories over an additional 10,000 generations. The parameters of these simulations are detailed in Sections 5.1 and 5.2, respectively.

## 4. Justification for assumptions

Here, we justify the assumptions that we relied upon in deriving the first two moments of change in allele frequency (see Section 2.1; modeling assumptions are motivated in the introduction of Chapter 2). We rely in part on self-consistency arguments, which should not be mistaken for being circular: specifically, we make assumptions about the behavior of the system and show that the solution to which we arrive satisfies these assumptions.

### 4.1. Normal and isotropic phenotypic distribution around the optimum

The assumption that the phenotypic distribution is well approximated by a normal distribution stems from the additive model of quantitative traits. By assuming that the phenotype arises from many additive contributions and that these additive contributions arise from some underlying distribution, normality follows from the law of large numbers. In terms of model parameters, we would expect normality to hold if the rate of mutations affecting the trait is sufficiently large, i.e., when $2NU \gg 1$.

We further assume that the phenotypic distribution is isotropic and its mean is at the optimum. Isotropy of the phenotypic distribution follows from assuming isotropy in the mutational input. In Section 5.4, we explore the consequences of anisotropy in the mutational input. In Section 4.4, we further show that the fluctuations of the mean phenotype around the optimum over time have negligible effects on allelic dynamics; a similar argument applies to fluctuations in the variance.

### 4.2. The phenotypic variance satisfies $\sigma^2 \ll w^2$

With the mean phenotype centered at the optimum, requiring that $\sigma^2 \ll w^2$ is equivalent to assuming that moving a standard deviation away from the mean phenotype entails only a minor reduction in fitness. This seems plausible for many phenotypes: if, for example, this assumption

did not hold for human height, then individuals whose height is a standard deviation or more away from the population mean would suffer a substantial reduction in fitness. Arguably, deviations from the mean height would then be recognized as a very common and severe disease.

Another line of argument that it is likely that $\sigma^2 \ll w^2$ is based on our results. If we assume that mutations are strongly selected, then our results suggest that

$$\sigma^2 = 2NU \cdot v_s = 4Uw^2/n. \quad \text{(A2.41)}$$

It follows that if the rate of mutations affecting the phenotype under consideration satisfies $U \ll 1$ then $\sigma^2 \ll w^2$. The number of mutations per diploid human genome per generation is estimated to be ~60 (5), and less than 10% of the genome is assumed to be functional (6), suggesting that the number of de novo mutations with any effect on function is less than 3 per haploid per generation. It then seems plausible that the (haploid) mutation rate affecting a specific trait satisfies $U \ll 1$. Assuming that mutations are weakly selected increases the variance in Eq. A2.41 only moderately and assuming the mutations are effectively neutral would suggest it is much smaller, leaving the above argument intact.

### 4.3. Mutational effect sizes satisfy $a^2 \ll w^2$

As we argued in the introduction of Chapter 2, variants for which the stronger condition $a^2 \ll \sigma^2$ holds account for most or all of the heritability explained in GWAS for many traits (e.g., 7, 8-10). Moreover, evidence for many traits suggests that the same is true for the variants that underlie the heritability that remains to be explained (11-14). Indeed, for this assumption to be violated, much of the genetic variance would have to arise from mutations that have a very large impact on fitness (i.e., with *s* on the order of 1). While this may be the case for some diseases (e.g., autism (15)), it does not appear to be the case for most quantitative traits that have been examined.

### 4.4. Deviations of the mean phenotype from the optimum can be neglected

In reality, the mean phenotype of the population fluctuates around the optimum. Here, we derive equations for the dynamic of the mean phenotype in order to estimate the magnitude and timescale of these fluctuations. We then show that these fluctuations have a negligible effect on the first two moments of change in allele frequency and thus on the results that follow from these moments.

We begin by deriving the first and second moment of change in mean phenotype. To this end, we assume the distribution of phenotypes is a multivariate normal centered around a mean phenotype, $\bar{r}$, i.e. that

$$\mathrm{f}(\vec{r}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{|\vec{r}-\bar{r}|^2}{2\sigma^2}\right). \quad \text{(A2.42)}$$

The expected change in mean phenotype due to selection in one generation is therefore

$$\mathrm{E}(\Delta\bar{r}) = \frac{\int_{\vec{r}} \mathrm{f}(\vec{r})\mathrm{W}(\vec{r})\vec{r}}{\int_{\vec{r}} \mathrm{f}(\vec{r})\mathrm{W}(\vec{r})} - \bar{r} = -\frac{\sigma^2}{w^2+\sigma^2}\bar{r} \approx -\frac{\sigma^2}{w^2}\bar{r}. \quad \text{(A2.43)}$$

By the same token, the variance in $\Delta\bar{r}$ is simply the sampling variance

$$\mathrm{V}(\Delta\bar{r}) \approx \frac{\sigma^2}{N}, \quad \text{(A2.44)}$$

where in both cases we relied on the assumption that $\sigma^2 \ll w^2$.

These two moments define an Ornstein-Uhlenbeck process in $\bar{r}$, allowing us to rely on well-known results (16). Notably, when the mean phenotype $\bar{r}$ starts far from the optimum, it decays exponentially to the optimum with exponent $\sigma^2/w^2$ (see Section 5.1 below). At steady state, $\bar{r}$ will fluctuate with mean zero and $\mathrm{E}(\bar{r}^2) = nw^2/2N$ over a time scale of $w^2/\sigma^2$ generations. The typical displacement of $\bar{r}$ in any given direction will be $\sqrt{w^2/2N}$, reflecting a balance between drift and the pull of selection toward the optimum.

Next we show that these fluctuations of the mean have negligible effects on allelic trajectories. To this end, we derive the first two moments of change in allele frequency, but this time, we include the effect of the displacement of $\bar{r}$ from the optimum. While the second moment remains the same, the first moment becomes

$$\mathrm{E}(\Delta q) \approx -\frac{\bar{r}\cdot\vec{a}}{w^2}pq - \frac{a^2}{w^2}pq\left(\frac{1}{2}-q\right) = -\frac{\bar{r}_a}{\sqrt{w^2/2N}}\frac{\sqrt{S}}{2N}pq - \frac{S}{2N}pq\left(\frac{1}{2}-q\right), \qquad \text{(A2.45)}$$

where $\bar{r}_a$ is $\bar{r}$'s component in the direction of $\vec{\square}$. However, our analysis establishes that $\frac{\bar{r}_a}{\sqrt{w^2/2N}}$ is a scalar on the order of 1, which fluctuates around zero on a timescale of $w^2/\sigma^2$. We can therefore compare the first term in the above equation, which represents directional selection, and the second term, which represents stabilizing selection. When stabilizing selection is strong, $S \gg 1$, the stabilizing selection term dominates over the directional selection term. In contrast, when selection is weak, i.e., $S \approx 1$ or smaller, then in any given generation, the directional term is not necessarily negligible. However, in this case, both terms affect substantial change in allele frequency only over a timescale of $2N$ generations; on this timescale, if $2N \gg w^2/\sigma^2$, the directional effect would average to zero. The directional term will become important only when $2N \leq w^2/\sigma^2$, that is $\sigma^2 \leq w^2/2N$. For $\sigma^2$ to be that small, virtually all alleles must have $S \ll 1$, such that their trajectories will be determined by drift, not selection. In summary, regardless of the selection acting on an allele, fluctuations of the mean phenotype around the optimum will have a negligible effect on its trajectory.

## 5. Model robustness

In this section, we consider the sensitivity of our results to relaxing some of the simplifying modeling assumptions about selection and mutation. Specifically, we show our results to be robust to moderate changes to the optimal phenotype; small asymmetry in the mutational input; the presence of major loci maintained at high frequency by selection on traits that are not included in the model; as well as to most forms of anisotropic mutation.

### 5.1. Changes to the optimal phenotype

We first consider how changes to the optimal phenotype over time would affect our results. It is easy to imagine how events such as migration from Africa to Europe or the onset of agriculture may have introduced rapid changes in optimal phenotypes. In order to evaluate the potential impact of such events, we consider how an instantaneous change to the optimal phenotype would affect the allelic dynamics. Similar models have recently been analyzed in the limit of infinite population size (17, 18).

We begin by considering how such an instantaneous change to the optimum would affect the mean phenotype. If the shift to the optimum is small, on the order of the fluctuations in the mean phenotype at steady state or smaller, then the arguments provided in Section 4.4 will still hold and the shift would have a negligibly small effect on our results. We therefore assume that the shift in optimum, $\vec{z}$, is large compared to the scale of fluctuations ($z^2 \gg w^2/2N$). This assumption means that we can use a deterministic approximation (based on Eq. A2.43) and describe the change in mean phenotype in a single generation by

$$\Delta\bar{r} \approx \mathrm{E}(\Delta\bar{r}) = -\frac{\sigma^2}{w^2}(\bar{r} - \vec{z}) \qquad \text{(A2.46)}$$

(neglecting higher moments). Further assuming that the mean phenotype was at the optimum, $\vec{0}$, before the optimum shifted (at time $t = 0$) and neglecting changes to the genetic variance $\sigma$, we find that

$$\bar{r}(t) = \vec{z}\left(1 - \exp\left(-\frac{\sigma^2}{w^2}t\right)\right). \quad \text{(A2.47)}$$

Thus, the mean $\bar{r}$ adapts to the new optimum on a timescale of $w^2/\sigma^2$ generations (see (19) for a similar derivation).

We can rely on this approximation to learn when a shift in optimum will have negligible effects on allele trajectories. Recalling Eq. A2.45, the first moment of change in allele frequency is given by

$$\mathrm{E}(\Delta q) \approx -\frac{(\bar{\mathrm{r}}(t)-\vec{z})\cdot\vec{a}}{w^2}pq - \frac{a^2}{w^2}pq\left(\frac{1}{2} - q\right), \quad \text{(A2.48)}$$

where, based on our approximation (Eq. A2.47), the directional selection term introduced by the shift in optimum takes the time-dependent form

$$\mathrm{E}(\Delta_D q) = -\frac{(\bar{\mathrm{r}}(t)-\vec{z})\cdot\vec{a}}{w^2}pq \approx \frac{\vec{z}\cdot\vec{a}}{w^2}\exp\left(-\frac{\sigma^2}{w^2}t\right)pq. \quad \text{(A2.49)}$$

The effect of this directional term over the entire adaptive trajectory can be quantified by comparing the expected allele frequency after adaptation to the shift, $q_D$, with initial frequency before the shift, $q_0$, i.e.,

$$\ln(q_D/q_0) = \int_t \frac{\mathrm{E}(\Delta_D q)}{q(t)} = \frac{\vec{z}\cdot\vec{a}}{w^2}\int_t p(t)\exp\left(-\frac{\sigma^2}{w^2}t\right) < \frac{\vec{z}\cdot\vec{a}}{w^2}\int_t \exp\left(-\frac{\sigma^2}{w^2}t\right) = \frac{\vec{z}\cdot\vec{a}}{\sigma^2}. \quad \text{(A2.50)}$$

This result suggests that the relative change in allele frequency will be negligible so long as

$$(\vec{z}\cdot\vec{a})/\sigma^2 \ll 1. \quad \text{(A2.51)}$$

This condition suggests that mutations with smaller effects would be less affected by the shift in the optimum. It further suggests that alleles that satisfy $a^2 \ll \sigma^2$, as appears to be the case for most loci discovered in GWAS (e.g., 7, 8-10), will be negligibly affected by shifts in optimum

on the order of the total genetic variation (i.e., $z \leq \sigma$). These analytic predictions are confirmed by simulations (Fig. A2.7).

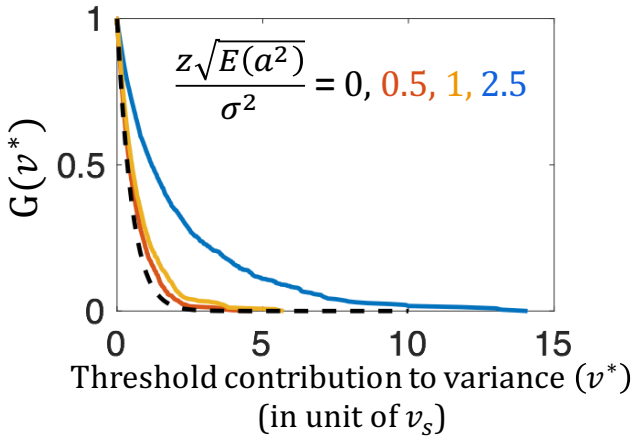



**Figure A2.7.** Distribution of the contributions of sites to variance after a shift in the optimum. The y-axis is the proportion of the variance explained by sites that contribute more than $v^*$ to the variance. The theoretical prediction without adaptation is shown in dashed black, and simulation results for different shifts in the optimal phenotype are shown in color. When the root mean square of $(\vec{z} \cdot \vec{a})/\sigma^2$ becomes larger than 1, directional selection substantially affects allele frequencies and therefore the contributions of sites to variance, as predicted by Eq. A2.51. (Since mutation is symmetric the mean of $(\vec{z} \cdot \vec{a})/\sigma^2$ is zero and we quantify its characteristic value by its root mean square $z\sqrt{E(a^2)}/\sigma^2$.) Simulations were run with an exponential distribution of selection coefficients with $\mathrm{E}(S) = 25$, $N = 1{,}000$, $n = 1$, $U = 0.01$, and a burn-in time of 10,000 generations. Results were taken 50 generations after the shift in optimum, which, for these parameters, is just after the population mean has reached the new optimum.

### 5.2. Asymmetric mutational input

In this section, we consider the sensitivity of our results to asymmetries in the mutational input, i.e., to the case in which mutations in a given direction in trait space are more likely to arise than mutations in the opposite direction (see (20) for treatment of this problem in the limit of high per-site mutation rate).

An asymmetric mutational input introduces a shift in the mean phenotype, $\bar{r}$, every generation. With new mutations arising at frequency $1/2N$, the expected shift is

$$\Delta_{\mathrm{M}}\bar{r} = 4NU\ \mathrm{E}_{\mathrm{M}}(\vec{a}) \cdot 1/2N = 2U\mathrm{E}_{\mathrm{M}}(\vec{a}), \quad \text{(A2.52)}$$

where $\mathrm{E}_{\mathrm{M}}$ is the expectation over newly arising mutations. For each trait, effects have a characteristic size $\sqrt{\mathrm{E}(a^2)/n}$=$\sqrt{v_s}\sqrt{\mathrm{E}(S/4)}$. The characteristic effect size sets the scale for the maximal shift in any direction, that is $\|\Delta_M\bar{r}\|$ is of the order of $2U\sqrt{\mathrm{E}(a^2)/n}$ or smaller. We therefore parameterize the shift in mean phenotype due to new mutations by

$$\Delta_{\mathrm{M}}\bar{r} = 2U\mathrm{E}_{\mathrm{M}}(\vec{a}) = 2U\sqrt{\mathrm{E}(a^2)/n}\,\vec{b}, \quad \text{(A2.53)}$$

where the vector $\vec{b}$ parameterizes the strength and direction of the bias and $b = \left\|\vec{b}\right\|$ is assumed to be << 1.

At steady state, the mutational shift must be offset by selection, such that

$$\Delta_{\mathrm{M}}\bar{r} + \Delta_{\mathrm{D}}\bar{r} + \Delta_{\mathrm{S}}\bar{r} = \vec{0}, \quad \text{(A2.54)}$$

where $\Delta_{\mathrm{D}}\bar{r}$ and $\Delta_{\mathrm{S}}\bar{r}$ are the expected shifts due to directional and stabilizing selection, respectively. We previously found that the expected directional shift is

$$\Delta_D\bar{r} = -\frac{\sigma^2}{w^2}\bar{r}, \quad \text{(A2.55)}$$

where $\bar{r}$ denotes the mean phenotype (see Eq. A2.43). As we show next, when mutations are strongly selected, stabilizing selection offsets the mutational shift to maintain the mean phenotype at the optimum, implying that directional selection is negligible. In contrast, when mutations are effectively neutral, stabilizing selection is negligible and a directional term might not be negligible by comparison. However, as long as asymmetry is small, $b \ll 1$, we show that this directional term is not large enough to change the allele dynamics, both when all mutations are effectively neutral and when some mutations are strongly selected.

First, we consider the shift in mean phenotype due to stabilizing selection. This shift arises because, with asymmetric mutational input, the distribution of phenotypes becomes skewed. Therefore, stabilizing selection may change the mean phenotype even if it is at the optimum. We have already shown (Eq. A2.15) that the expected change in allele frequencies per generation due to stabilizing selection at any given site *i* is

$$E(\Delta q_i) = -\frac{a_i^2}{w^2} p_i q_i \left(\frac{1}{2} - q_i\right). \quad \text{(A2.56)}$$

The expected change in mean phenotype can then be calculated by adding up the contributions over sites

$$\Delta_S \bar{r} = -\mathrm{E}\left(\sum_i 2\vec{a}_i \frac{a_i^2}{w^2} p_i q_i \left(\frac{1}{2} - q_i\right)\right). \quad \text{(A2.57)}$$

The right-hand side of this equation reflects the skewness of the phenotypic distribution. Indeed, in one dimension, it can be shown that

$$\Delta_S \bar{r} = -\frac{\mu_3(r)}{2w^2}, \quad \text{(A2.58)}$$

with $\mu_3(r)$ being the third central moment of the phenotypic distribution. In *n*-dimensions, for every direction $x$,

$$\Delta_S \bar{r}_x = -\frac{1}{2w^2}\mathrm{E}((\vec{r} - \bar{r})_x(\vec{r} - \bar{r})^2) = -\frac{1}{2w^2}\sum_k \mu_3(\vec{r})_{xkk}, \quad \text{(A2.59)}$$

with $\mu_3(\vec{r})_{jkl} = \mathrm{E}\big((\vec{r} - \bar{r})_j(\vec{r} - \bar{r})_k(\vec{r} - \bar{r})_l\big)$.

When sites are under strong selection, $\Delta_S \bar{r}$ takes a simple form. Assuming the asymmetry is small, the shift due to stabilizing selection can be expanded in orders of $b$. The leading term in the frequency distribution takes the same form as it does without the bias. For strongly selected alleles with no bias, $q \ll 1$ and therefore the frequency dependence in this term can be

approximated by $pq\left(\frac{1}{2}-q\right)\approx\frac{1}{2}q$. Moreover, $q$ scales with $1/a^2$, implying that the distribution of $a^2q$ is independent of $\vec{a}$ and that $\mathrm{E}(a^2q)=w^2/N$ (see Section 3.1).

Therefore, when all sites are strongly selected, the leading term in the shift due to stabilizing selection is

$$\Delta_S^0\bar{r}=-\mathrm{E}\left(\sum_i 2\vec{a}_i\frac{a_i^2}{w^2}\frac{q_i}{2}\right)=-\frac{\mathrm{E}(a^2q)}{w^2}\mathrm{E}(\textstyle\sum_i\vec{a}_i)=-\frac{1}{N}2NU\mathrm{E}_M(\vec{a})=-\Delta_M\bar{r}. \qquad \text{(A2.60)}$$

Thus, to a first order in $b$, the shift of the mean phenotype due to stabilizing selection offsets the mutational shift, implying that there will be no directional term and that the allele dynamics will not be affected by asymmetry.

When alleles are instead effectively neutral, then $a^2/w^2\ll 1/2N$ (see Section 2.2) and allele frequencies are well approximated by the neutral sojourn time, $\tau(q)\approx 2/q$. The shift due to stabilizing selection then satisfies

$$\Delta_S\bar{r}=-\mathrm{E}\left(\sum_i 2\vec{a}_i\frac{a_i^2}{w^2}p_iq_i\left(\frac{1}{2}-q_i\right)\right)\approx-\mathrm{E}\left(pq\left(\frac{1}{2}-q\right)\right)\mathrm{E}\left(\sum_i 2\vec{a}_i\frac{a_i^2}{w^2}\right)$$

$$=-\frac{1}{6}\,\mathrm{E}\left(\sum_i 2\vec{a}_i\frac{a_i^2}{w^2}\right)\ll-\frac{1}{N}\mathrm{E}(\textstyle\sum_i\vec{a}_i)=-\Delta_M\bar{r}, \qquad \text{(A2.61)}$$

implying that it makes a negligible contribution to offsetting the mutational shift. In this case, the mutational effect on the mean phenotype is therefore offset by directional selection, where

$$\Delta_D\bar{r}=-\frac{\sigma^2}{w^2}\bar{r}\approx-\Delta_M\bar{r}, \qquad \text{(A2.62)}$$

indicating a displacement of the mean phenotype from the optimum

$$\bar{r}=\frac{\square^2}{\sigma^2}\Delta_M\bar{r}. \qquad \text{(A2.63)}$$

This displacement introduces a directional selection term into the first moment of change in allele frequency that, if large enough, could alter allele dynamics (see Section 4.4).

However, when all alleles are effectively neutral, we have

$$\bar{r} = \frac{w^2}{\sigma^2}\Delta_M \bar{r} = \frac{w^2}{2NUv_s \mathrm{E}(S)/2} 2U\sqrt{v_s}\sqrt{\mathrm{E}(S/4)}\,\vec{b} = \frac{1}{2N}\frac{2w^2}{\sqrt{v_s}\sqrt{\mathrm{E}(S)}}\,\vec{b}, \quad \text{(A2.64)}$$

and therefore the scaled directional selection coefficient, for an allele with effect size $\vec{a}$ and scaled stabilizing selection coefficient $S = 2N\frac{a^2}{w^2}$, is of the order of

$$2N\frac{\bar{r}\cdot\vec{a}}{w^2} = \frac{2b_a a}{\sqrt{v_s}\sqrt{\mathrm{E}(S)}} \sim \frac{2(b/\sqrt{n})a}{\sqrt{v_s}\sqrt{\mathrm{E}(S)}} = \frac{b\sqrt{v_s}\sqrt{S}}{\sqrt{v_s}\sqrt{\mathrm{E}(S)}} = \sqrt{\frac{S}{\mathrm{E}(S)}}\,b, \quad \text{(A2.65)}$$

with $b_a \sim b/\sqrt{n}$ being the projection of $\vec{b}$ in the direction of $\vec{a}$. Since $b \ll 1$, for all alleles other than those with unusually large selection coefficients, the scaled directional selection coefficient will be much smaller than 1 and the trajectories will still be determined by drift and not selection. Even in this case, therefore, we do not expect asymmetry to affect allele dynamics.

Next, we consider the case where there is a mix of effectively neutral and strongly selected mutations. The existence of strongly selected mutations in addition to effectively neutral ones reduces the deviation of the mean phenotype from the optimum. Denoting the proportion of strongly selected mutations by $p_s$, we have

$$\bar{r} = \frac{w^2}{\sigma^2}U(1-p_s)\sqrt{v_s}\sqrt{\mathrm{E}(S^{e.n.})}\,\vec{b}, \quad \text{(A2.66)}$$

where $\mathrm{E}(S^{\mathrm{e.n.}}) \leq 1$ is the mean scaled stabilizing selection coefficient for effectively neutral mutations. Since $\sigma^2 > 2NUp_s v_s$, we can then obtain an upper bound to the magnitude of scaled directional selection coefficient for an allele with effect size $\vec{a}$ and scaled stabilizing selection coefficient $S = 2N\frac{a^2}{w^2}$:

$$\begin{aligned} 2N\frac{\bar{r}\cdot\vec{a}}{w^2} &= 2N\frac{1}{\sigma^2}U(1-p_s)\sqrt{v_s}\sqrt{\mathrm{E}(S^{e.n.})}\,\vec{b}\cdot\vec{a} \\ &< \frac{1}{Up_s v_s}U(1-p_s)\sqrt{v_s}\sqrt{E(S^{e.n.})}\,\vec{b}\cdot\vec{a} \sim \frac{1-p_s}{2p_s}\sqrt{E(S^{e.n.})}\,\sqrt{S}\,b. \end{aligned} \quad \text{(A2.67)}$$

With a substantial proportion of strongly selected sites, $(1-p_s)/2p_s$ is of the order of 1, and therefore $\frac{1-p_s}{2p_s}\sqrt{\mathrm{E}(S^{e.n.})}\,b \ll 1$. This condition implies that for effectively neutral alleles (i.e., $S \leq 1$), the scaled directional selection coefficient is $\ll 1$ and allele trajectories will be determined by genetic drift, whereas for strongly selected alleles (i.e., when $S \gg 1$), the scaled directional selection coefficient is $\ll S$ and therefore negligible compared to the scaled stabilizing selection coefficient.

Weakly selected alleles (with $1 < S < 30$) behave largely like strongly selected alleles except that stabilizing selection on them only partially cancels out the mutational bias (for example, for $S = 10$ only 85% of the bias is canceled). The rest of the bias is canceled by directional selection and therefore induces a small shift in the mean phenotype. It is straightforward to repeat the arguments given above and show that the shift in the mean phenotype for a trait with only weakly selected alleles or a mixture that includes weakly selected alleles negligibly affects allele trajectories.

Thus, we conclude that small asymmetry in mutation will not affect the allelic dynamic (see Fig. A2.8).

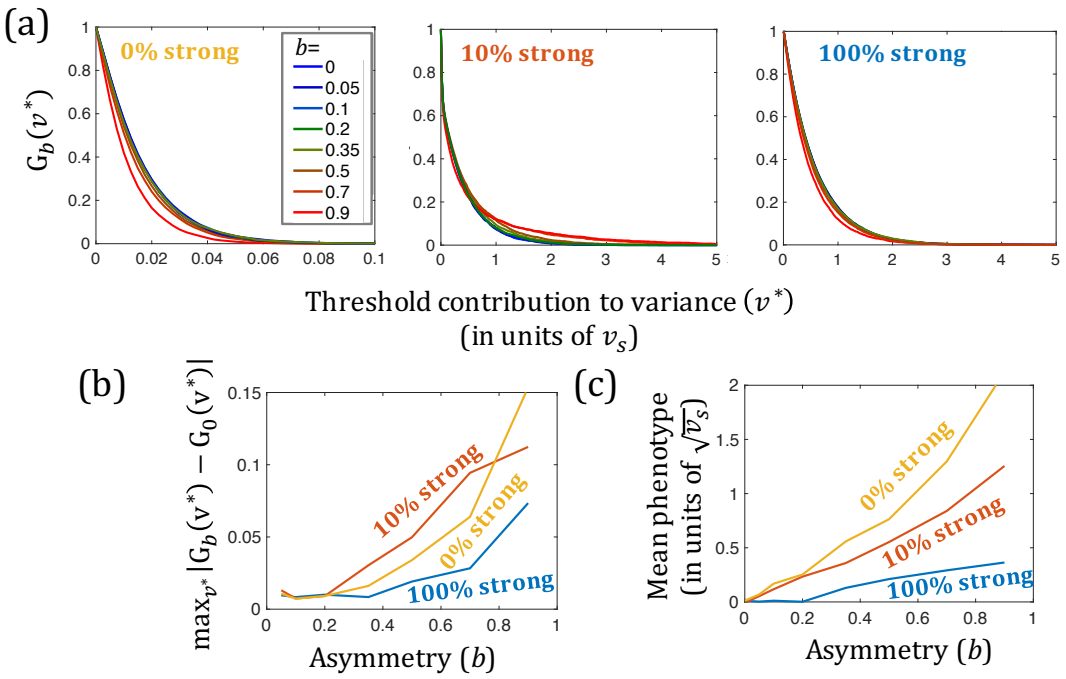



**Figure A2.8.** The effect of asymmetric mutational input on the contribution of sites to variance and the mean phenotype. (a) Proportion of genetic variance as a function of the threshold contribution to variance $v^*$, i.e., $G_b(v^*)$, for different bias strengths. (b) The maximal distance of $G_b(v^*)$ from $G_0(v^*)$, i.e. $\max_{v^*}|G_b(v^*) - G_0(v^*)|$, as a function of $b$. (c) The mean phenotype $\bar{r}$, in units of $\sqrt{v_s}$, as a function of mutational bias $b$. Simulations were run with $N = 1{,}000$, $n = 1$ and with different mixtures of effectively neutral (exponentially distributed with $E(S) = 0.1$) and strong (exponentially distributed with $E(S) = 50$) selection coefficients. Asymmetry was simulated by having more trait increasing than trait decreasing mutations; if $\beta$ is the proportion of trait increasing mutations then the asymmetry coefficient is $b = 2\beta - 1$. As expected, for small biases (when $b \ll 1$), there are no substantial changes in the distribution of the contribution of sites to variance. Simulations were run with a 10,000 generations burn-in period without asymmetry and then 10,000 generations with asymmetry and averaged over many runs (>300), with the number of runs varied across plots keep errors in (a) below 1%.

### 5.3. Major effect loci

In this section, we show that our results are insensitive to the presence of major loci, i.e., individual loci that contribute substantially to quantitative genetic variation. We have in mind, for example, loci whose alleles are maintained at high frequencies by balancing selection on a Mendelian trait but have pleiotropic effects on the quantitative traits under consideration (e.g.,

HLA loci (21, 22)). While such loci violate our assumptions, we show that they do not affect the dynamics at other loci that fulfill them.

To this end, we calculate the first two moments of change in allele frequency in the presence of a major locus. We denote the frequency and effect size of the focal allele by $q$ and $\vec{a}$, and the frequency and effect size of the major allele by $q_M$ and $\vec{a}_M$, respectively. As in our previous derivations (Section 2.1), the distribution of background phenotypic contribution from all other loci, $\vec{R}$, is well approximated by the normal distribution

$$\mathrm{f}\left(\vec{R}\middle|\vec{a}_M, q_M, \vec{a}, q\right) = \frac{1}{\left(2\pi(\sigma^2-\sigma_M^2)\right)^{n/2}} \exp\left(-\frac{\left(\vec{R}+2q\vec{a}+2q_M\vec{a}_M\right)^2}{2(\sigma^2-\sigma_M^2)}\right), \quad \text{(A2.68)}$$

where $\sigma_M^2$ is the contribution to genetic variance from the major locus. The population mean remains close to the optimum because any shift caused by the major locus is quickly compensated for by the other loci (see Section 4.4). We then average over both this distribution and the three genotypes at the major locus to calculate the mean fitness associated with each genotype at the focal locus. Namely,

$$W_{00} = (1-q_M)^2 \int_{\vec{R}} \mathrm{f}\left(\vec{R}\middle|\vec{a}_M, q_M, \vec{a}, q\right)\mathrm{W}\left(\vec{R}\right) + 2q_M(1-q_M) \int_{\vec{R}} \mathrm{f}\left(\vec{R}\middle|\vec{a}_M, q_M, \vec{a}, q\right)\mathrm{W}\left(\vec{R}+\vec{a}_M\right)$$
$$+q_M^2 \int_{\vec{R}} \mathrm{f}\left(\vec{R}\middle|\vec{a}_M, q_M, \vec{a}, q\right)\mathrm{W}\left(\vec{R}+2\vec{a}_M\right), \quad \text{(A2.69)}$$

and similarly for the other genotypes. In this way, we obtain the first moment of the change in allele frequency

$$\mathrm{E}(\Delta q) = -pq\frac{p(W_{00}-W_{01})+q(W_{01}-W_{11})}{\bar{W}} \approx -\frac{a^2}{w^2}pq\left(q-\frac{1}{2}\right), \quad \text{(A2.70)}$$

which is the same as we derived in the absence of a major locus (Eq. A2.15). Similarly, we find the second moment to be unaffected.

### 5.4. Anisotropic mutation

In this section, we consider how relaxing the assumption that the distribution of newly arising mutations is isotropic in trait space would affect our results. As noted, we can always choose an orthonormal coordinate system centered at the optimum, in which the trait under consideration varies along the first coordinate and a unit change in other traits (i.e., in other coordinates) near the optimum have the same effect on fitness. There is, however, no obvious reason for the distribution of newly arising mutations to be isotropic in this coordinate system (see (23) for generalizations of Fisher's Geometric Model along similar lines).

Anisotropy in mutation does not affect the moments of change in allele frequency, as these depend only on the selection on an allele or equivalently on its effect size but not on its direction in trait space. Anisotropy could affect the distribution of allelic effect sizes on the focal trait conditional on the selection acting on them. Here, we provide heuristic arguments suggesting that, barring extreme cases, we can define an effective number of traits $n_e$ and an effective strength of selection $w_e^2$ for which the relationship between selection and effect size in anisotropic models is well approximated by the relationship found for isotropic ones (Eqs. 9 & 11, which appear in the Results section of Chapter 2; Section 1.2).

We focus on a family of anisotropic mutational distributions that can be described as a projection of a multivariate normal distribution on the unit sphere in trait space. Namely, we draw the size of a mutation $a = \|\vec{a}\|$ from some distribution and to obtain its direction, we draw a vector $\vec{\alpha}$ from a multi-variate normal distribution $\mathrm{MVN}(0, \boldsymbol{\Sigma})$ and normalize it, i.e.,

$$\vec{a} = a\frac{\vec{\alpha}}{\alpha}, \tag{A2.71}$$

and therefore

$$a_1 = a\frac{\alpha_1}{\alpha}. \tag{A2.72}$$

This family of mutational distributions gives us a mathematically tractable framework with which to examine the behavior of our model with anisotropy.

With anisotropy, the behavior of our model greatly depends on the relative contribution of the focal trait to selection, which we parameterize by

$$\gamma_1 \equiv \frac{\mathrm{E}(\alpha_1^2)}{\mathrm{E}(\alpha^2)} = \frac{\boldsymbol{\Sigma}_{11}}{\mathrm{tr}(\boldsymbol{\Sigma})}. \tag{A2.73}$$

When selection acts mainly on our focal trait, i.e. when $\gamma_1 \approx 1$, then $|\alpha_1| \approx \alpha$ and therefore $a_1 \approx \pm a$. Such a relationship between the strength of selection and effect size is well approximated by an isotropic model with $n_e = 1$. We therefore focus on cases in which there is a significant pleiotropic contribution to selection, i.e., $\gamma_1$ is substantially less than 1. Anisotropy then has two effects: the first is to introduce heterogeneity in the strength of selection on different traits and the second is to introduce correlations in the effects of a mutation on different traits, notably between the focal trait and others.

We first consider the case in which the strength of selection differs among traits, but traits are uncorrelated, corresponding to a diagonal covariance matrix, $\boldsymbol{\Sigma}$. When many traits have a non-negligible contribution to selection, $\alpha^2 = \alpha_1^2 + \alpha_2^2 + \cdots \alpha_n^2$ would have a small coefficient of variation, i.e., $\mathrm{C_V^2}(\alpha^2) = \mathrm{V}(\alpha^2)/\mathrm{E}(\alpha^2)^2 \ll 1$, because of the law of large numbers. In this case,

$$a_1 = a\frac{\alpha_1}{|\vec{\alpha}|} = a\frac{\alpha_1}{\sqrt{E(\alpha^2)}}\left(1 + \mathrm{O}\left(\mathrm{C_V^2}(\alpha^2)\right)\right) \approx a\frac{\alpha_1}{\sqrt{\mathrm{E}(\alpha^2)}} = \frac{a}{\sqrt{1/\gamma_1}}\frac{\alpha_1}{\sqrt{\mathrm{E}(\alpha_1^2)}}, \tag{A2.74}$$

Since $\alpha_1/\sqrt{\mathrm{E}(\alpha_1^2)} \sim \mathrm{N}(0,1)$ and $s = \frac{1}{w^2}a^2$, this implies that, conditional on the selection coefficient $S$, the effect size on the focal trait will be distributed as

$$a_1 \sim \mathrm{N}\left(0, \frac{w^2}{n_e}s\right) \tag{A2.75}$$

with $n_e = 1/\gamma_1$. This is the same relationship between selection and effect size as in the high pleiotropy isotropic model with $n = n_e$ (Eq. 11, which appears in the Results section of Chapter 2). This result suggests the concept of an effective number of traits, which can be thought of as the number of traits that have the same effect on fitness as the focal one and are required to produce the same strength of selection on alleles. The effective number of traits describes the distribution of effect sizes both in the limit of high pleiotropy $n_e \gg 1$ and low pleiotropy $n_e \approx 1$ and simulations show that it describes the distribution, at least qualitatively, also for intermediate values of $n_e$ (Fig. A2.9).

However, there is an extreme scenario in which an effective number of traits cannot describe the distribution of effect sizes. This happens when $\mathrm{C_V^2}(\alpha^2) \geq 1$, that is when selection acts mainly on a small number of traits but our focal trait contributes very little to selection ($\gamma_1 \ll 1$). In this case, we might be tempted to use $n_e = 1/\gamma_1 \gg 1$ but, as Eq. A2.74 suggests, the high pleiotropy limit would be inadequate. In fact, the variance in selection on newly-arising mutations (due to the contribution of the selected traits) will result in a long-tailed distribution of effect sizes on the focal trait, which is not well-approximated by any isotropic model. Excluding these extreme cases, isotropic models provide a good approximation for the relationship between selection and effect size, even when there is heterogeneity in the strength of selection on different traits.

To illustrate the effect of heterogeneity in the strength of selection among traits, we consider a simple example in which all non-focal traits make the same contribution to selection and therefore can be modeled by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \\ \mathbf{0} & \boldsymbol{c^2} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{c^2} & \\ & \vdots & & \ddots \end{pmatrix}, \tag{A2.76}$$

where $c^2$ is the ratio between the expected fitness effects of non-focal and focal traits. In Fig. A2.9 we compare numerical results of this model with isotropic models with $n_e = 1/\gamma_1 = 1 + (n-1)c^2$.

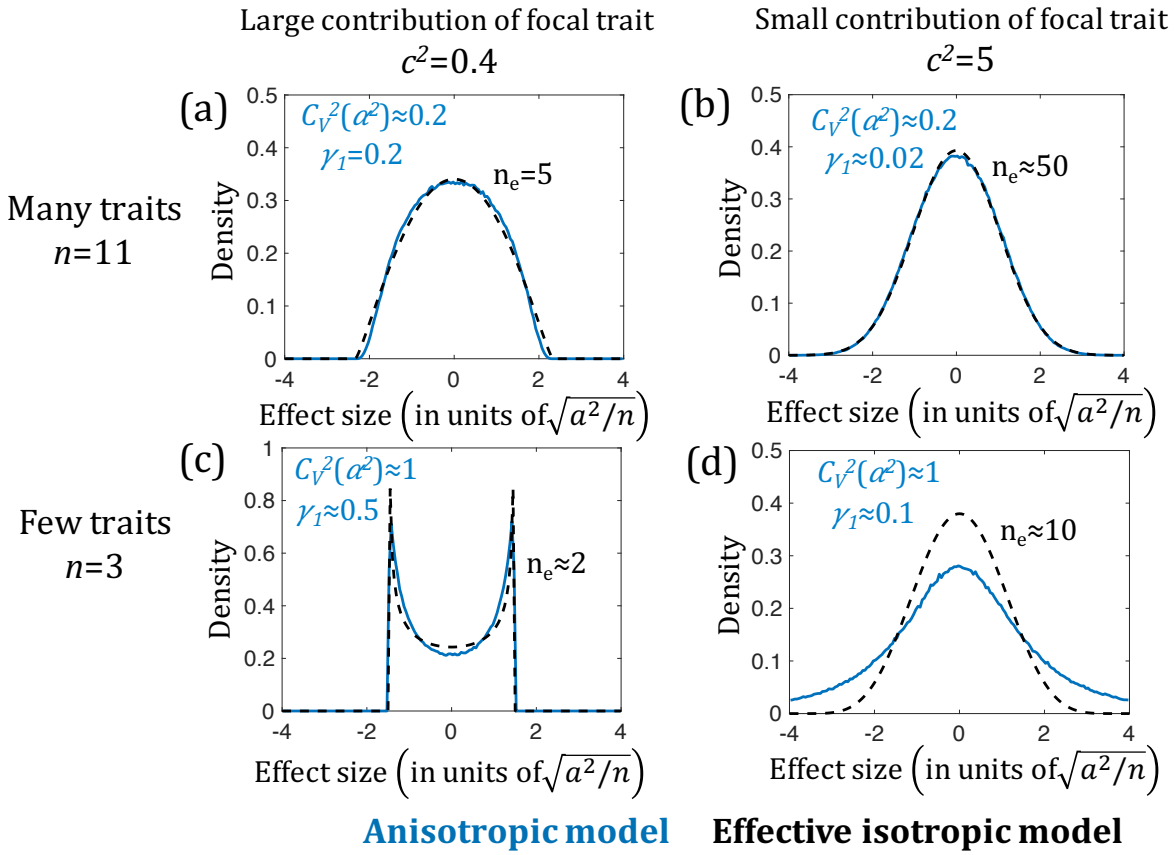



**Figure A2.9.** The effects of heterogeneity in the strength of selection on different traits on the distribution of effect sizes in the focal trait. Numerical results for models with the correlation matrix defined in Eq. A2.76 are shown in blue and the corresponding isotropic model in black dashes. When there are many selected traits, an isotropic model with $n_e = 1/\gamma_1 = 1 + (n-1)c^2$ provides a good approximation of the distribution of effect sizes, both when the focal trait contributes substantially to selection (a) and when it does not (b). When there are few traits, an isotropic model with $n_e = 1/\gamma_1 = 1 + (n-1)c^2$ provides a good approximation only when the focal trait contributes substantially to selection (c & d).

Next, we consider the case in which the effect sizes on different traits are correlated, i.e., when the covariance matrix $\boldsymbol{\Sigma}$ has off-diagonal terms. $a_1^2 \propto \alpha_1^2/\alpha^2$ and therefore we parameterize the effect of these terms using the correlation between $\alpha_1^2$ and $\alpha^2$, $\rho^2 \equiv \mathrm{corr}(\alpha^2, \alpha_1^2)$. If the

correlation is small, $\rho^2 \ll 1$, then our previous reasoning holds. In the other extreme, when all selected traits are highly correlated with the focal trait, i.e. $\rho^2 \approx 1$, then the proportional contribution of the focal trait to selection is constant, $\alpha_1^2/\alpha^2 \approx \mathrm{E}(\alpha_1^2)/\mathrm{E}(\alpha^2) = \gamma_1$, and the effect size is $a_1 = \pm\sqrt{\gamma_1}\, a$. This model is therefore equivalent to an isotropic one with $n_e = 1$ and $w_e^2 = \gamma_1 w^2$; the latter change corresponds to increasing the strength of selection on the focal trait to account for selection on the other, highly-correlated traits. Intermediate cases are more complex: while effect sizes are still of the order of $\sqrt{\gamma_1}\, a$, the shape of the distribution of effect sizes is intermediate between the single trait and high pleiotropy limits. Isotropic models with an effective number of traits, $n_e < 1/\gamma_1$, and increased selection $w_e^2 = \gamma_1 n_e w^2$ can describe these cases qualitatively but may not completely capture the distribution of effect sizes. The value of $n_e$ would change from 1 when $\rho^2 \to 1$ to $1/\gamma_1$ when $\rho^2 \to 0$. Note that with a large number of traits, very strong correlations among many of the traits will be necessary in order to create a large enough $\rho^2$ to have a significant effect on $n_e$ (see Fig. A2.10).

To illustrate the effect of correlations among traits, we consider the following simple example (Fig. A2.10). We assume the correlation matrix $\boldsymbol{\Sigma}$ takes the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{1} & \boldsymbol{r^2} & \boldsymbol{r^2} & \\ \boldsymbol{r^2} & \mathbf{1} & \boldsymbol{r^2} & \cdots \\ \boldsymbol{r^2} & \boldsymbol{r^2} & \mathbf{1} & \\ & \vdots & & \ddots \end{pmatrix}, \tag{A2.77}$$

meaning that that all traits contribute equally to the fitness and every pair of traits has the same correlation coefficient $r^2$. When $r^2 = 0$ this becomes an isotropic model. When $r^2 = 1$, effect sizes are always identical for every trait; thus, this case is equivalent to having only one trait with the strength of selection increased $n$-fold. Intermediate cases can be approximated by finding an effective number of traits $n_e < 1/\gamma_1 = n$, such that an isotropic model with $n_e$ and $w_e^2 =$

$\gamma_1 n_e w^2 = w^2 n_e/n$ qualitatively describes the distribution of effect sizes. Numerical results of this model are shown in Fig. A2.10.

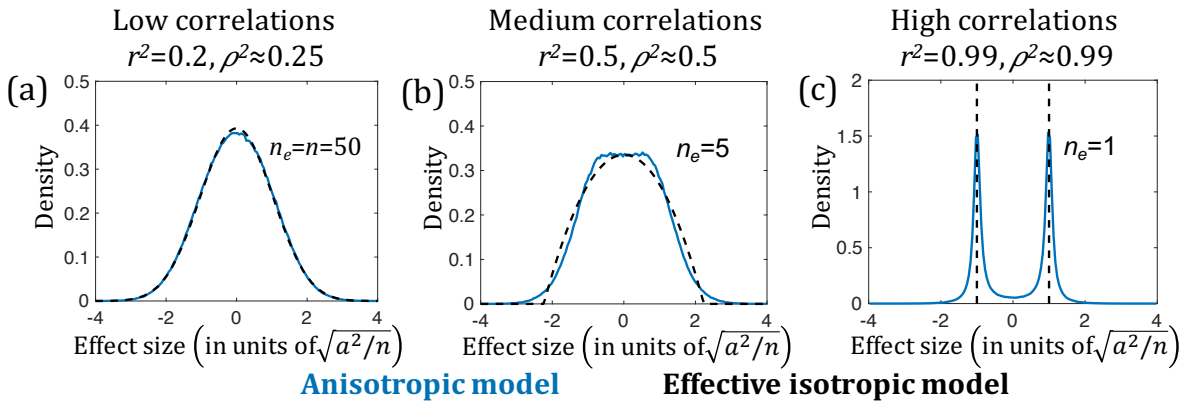



**Figure A2.10.** Effects of correlations among traits on the distribution of effect sizes. Numerical results for our model with the correlation matrix defined in Eq. A2.77 and $n = 50$ traits are shown in blue and the corresponding isotropic model in black dashes. (a) When correlations are low, the isotropic model approximates the distribution of effect sizes well. (b) With large correlations, we need to use an effective number of traits, in this example $n_e = 5$, and rescale selection, in this case to $w_e^2 = w^2 n_e/n = w^2/10$, in order to approximate the distribution of effect sizes. (c) When the correlations approach 1, the distribution of effect sizes becomes singular and approaches the distribution for an isotropic model with $n_e = 1$ and $w_e^2 = w^2/n = w^2/50$.

## 6. The power to detect loci in GWAS

In this section, we summarize the results that we rely on in connecting our theoretical predictions with the observations from GWAS (see Discussion in Chapter 2). These results provide a first approximation of the power to detect loci in GWAS in re-sequencing and genotyping studies. In Section 6.3 we consider potential complications that arise when GWAS do not identify causal loci but rather SNPs in LD with them.

### 6.1. Re-sequencing studies

First, we consider how the power to identify a locus in a GWAS depends on its contribution to genetic variance. To this end, we follow Sham and Purcell (24) in assuming a simplified model for a GWAS in which loci are detected using a linear regression of the phenotype against the genotype at individual loci, and the dependence of phenotype on genotype follows an additive model. The slope of the regression (the regression coefficient), which is also the estimate of the effect size, $\hat{a}_1$ is then approximately normally distributed as

$$\hat{a}_1 \sim \mathrm{N}\left(a_1, \frac{V_P/m}{2x(1-x)}\right), \quad \text{(A2.78)}$$

where $a_1$ is the true effect size and $x$ is the minor allele frequency at the locus (which, due to the large study sizes, we assume to be estimated without error), $V_P$ is the total phenotypic variance, and $m$ the study size (which in reality may be an effective size reflecting study design, e.g., when the sample is split into discovery and validation panels) (24).

Under the null hypothesis, the true effect size is 0, meaning that

$$\hat{a}_{1_{\mathrm{null}}} \sim \mathrm{N}\left(0, \frac{V_P/m}{2x(1-x)}\right) \quad \text{(A2.79)}$$

and therefore, the estimated contribution to variance has a chi-squared distribution with one degree of freedom

$$\frac{\hat{v}_{\text{null}}}{V_P/m} = \frac{2\hat{a}^2_{1\text{null}}x(1-x)}{V_P/m} \sim \chi^2_1. \quad \text{(A2.80)}$$

The power to identify a locus as significant with p-value $p^*$ is the probability that the estimated contribution of the locus to variance, $\hat{v}$, is large enough that

$$\Pr(\hat{v}_{\text{null}} > \hat{v}) < p^*. \quad \text{(A2.81)}$$

This condition can be translated into a threshold contribution to variance $v^*$ for which loci with $\hat{v} > v^*$ are considered significant, i.e. $\Pr(\hat{v}_{\text{null}} > v^*) = p^*$, with $v^*$ given by

$$\frac{v^*}{V_P/m} = 2\left(\text{erf}^{-1}(1-p^*)\right)^2, \quad \text{(A2.82)}$$

and erf denoting the error function. The power to identify a locus with a contribution $v$ to the genetic variance is $\Pr(\hat{v} > v^*|v)$ and the distribution of $\hat{v}$ is given by

$$\frac{\hat{v}}{V_P/m} = \frac{2\hat{a}^2x(1-x)}{V_P/m} \sim \chi^2_1\left(\frac{v}{V_P/m}\right), \quad \text{(A2.83)}$$

where $\chi^2_1$ denotes a non-central chi-squared distribution with one degree of freedom. Therefore, power is given by

$$\text{H}(v, p^*) = \Pr(\hat{v} > v^*|v) = \text{h}_+\left(\frac{v}{V_P/m}, p^*\right) + \text{h}_-\left(\frac{v}{V_P/m}, p^*\right), \quad \text{(A2.84)}$$

where $\text{h}_\pm(y, p^*) = \frac{1}{2}\left(1 \pm \text{erf}\left(\sqrt{y/2} \mp \text{erf}^{-1}(1-p^*)\right)\right)$ and the two terms correspond to the estimated and true effect sizes having the same or opposite sign.

The form of the power function carries important implications (Eq. A2.84 and Fig. A2.11). Notably, it shows that (in this approximation) power depends only on the contribution of a locus to variance, and this contribution should be measured relative to, or in units of, $V_P/m$. This scale makes intuitive sense, because the total phenotypic variance generates the background noise for detecting an individual locus, and the background noise is inverse proportional to the study size.

In particular, the threshold contribution to variance $v^*$, as defined above, is proportional to $V_P/m$ and is also the contribution to variance at which power is 50%, i.e.,

$$\mathrm{H}(v^*, p^*) = 1/2. \tag{A2.85}$$

The power function can then be approximated by a step function (see Fig. A2.11)

$$\mathrm{H}(v) \approx \Theta(v - v^*) = \begin{cases} 1 & v > v^* \\ 0 & v < v^* \end{cases}. \tag{A2.86}$$

This will be a good approximation when the number of loci that fall at intermediate range (e.g., with power between 0.1 and 0.9) is negligible compared to the number that falls outside this range.

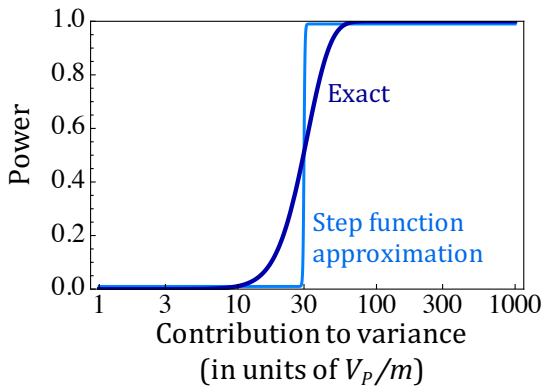



**Figure A2.11.** The power to detect loci as a function of their contribution to genetic variance (given in units of $V_P/m$). Shown are the exact power function (Eq. A2.84) and its step function approximation (Eq. A2.86) for $p = 5 \cdot 10^{-8}$.

Further insights come from considering this power function in conjunction with our theoretical results (Section 3). Notably, our results suggest that the first loci to be detected, those that contribute the most to variance, are intermediate and strongly selected, and that their contributions to variance are on the order of $v_s$. We therefore expect GWAS to begin to identify loci (and account for substantial genetic variance) when $v^*$ is on the order of $v_s$, i.e., when the study size $m$ is on the order of $V_P/v_s$. We would further expect the rate of increase in identifying new loci (and in accounting for genetic variance) to be similar for different traits when variance is measured in units of $v_s$.

## 6.2. Genotyping

Most current GWAS rely on genotyping instead of re-sequencing, resulting in an additional loss of power (26). Specifically, these studies impute the alleles at loci that are not included in the genotyping platform (27), and the imputation becomes imprecise when the imputed alleles are rare (Fig. A2.12). If causal loci with rare minor alleles are included in GWAS, this imprecision leads to an under-estimation of their effect size, resulting in loss of power (26). For loci with MAF $x$ and effect size $a$, the expected estimate of the effect size would be reduced by a factor of $r(x)$, where $r^2(x)$ is the mean correlation between the imputed and real alleles (28), and the distribution of estimates can be approximated by

$$\hat{a}_1 \sim \mathrm{N}\left(r a_1, \frac{V_P/m}{2x(1-x)}\right). \qquad \text{(A2.87)}$$

Employing the reasoning of the previous subsection, we can therefore approximate the power to detect a locus by $\mathrm{H}(r^2 v, p^*)$, where H is the power function defined in Eq. A2.84.

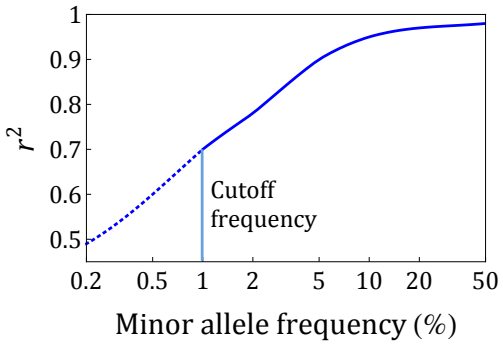



**Figure A2.12**. The precision of imputation decreases with MAF. Specifically we show the mean correlation between imputed and real genotypes as function of minor allele frequency, for a study using an Illumina 1M SNP array and the 1000 genomes phase III as an imputation panel (based on Extended Fig. 9A in (29)). We approximate the effect on power by excluding loci with MAF < 1% and assuming that loci with greater MAFs are imputed correctly.

In practice, GWAS often include only loci with MAF above a threshold, which is chosen to ensure precise imputation. We therefore approximate the effect of genotyping on power by

excluding loci below a threshold MAF and assume that loci that exceed this threshold are imputed correctly.

### 6.3. Tagging

Our inference is predicated on the assumption that the distribution of estimated variances among genome-wide significant (GWS) associations faithfully reflects the distribution among causal loci. We have no obvious alternative but to make this assumption, and arguably, the good fit of our theoretical predictions to the distribution of variances among associations provides some support for this assumption. While this assumption cannot be directly tested at present, existing arguments and evidence suggest that it is plausible, for reasons that we briefly review.

Most of the variants discovered by GWAS are common. Specifically, all but one of the GWS associations for height and BMI, which we rely upon in our inference, have MAF>1%, and the MAF of most associations is considerably greater. In considering the validity of our assumption, we therefore consider what could be tagged by such common associations. One possibility is that a given common association is tagging a single common, causal variant. Given the accuracy of imputation for common variants (see Fig. A2.12), we would therefore expect that the tagging variant would be in almost perfect LD with the causal one (including the possibility that the association is actually with the causal variant). If that were the case, then we would expect the estimated frequency and effect size, and thus the estimated contribution to genetic variance, to be very similar to those of the causal variant. A second possibility is that a given association tags several common causal variants within the same genomic region. The number of causal variants would likely be small, as otherwise the tagging allele is highly unlikely to be in LD with causal alleles that affect the trait in the same direction. If that were the case, given the accuracy of

imputation of the causal alleles, we would expect conditional analysis (e.g., 30) to successfully distinguish between the different causal variants, thus returning us to the previous scenario.

A third possibility involves a common association tagging rare, causal variants (25). While a single rare, causal variant would have to have an unreasonably large effect size in order to result in a common GWS association (31), it has been argued that several rare, causal variants in the same genomic region may be tagged by a single "synthetic association" (25). In this case, the relatively low LD between the association and each of the causal variants would imply that the estimated contribution to variance of the association would have to be much smaller than the combined contribution of the causal variants (25, 31). If this were the case for many associations identified in GWAS, it would violate the premise of our inference.

However, multiple lines of evidence suggest that it is not a common occurrence. One is that, where data is available, associations often replicate across populations. For example, there is considerable overlap between GWS associations for height in Europeans and East-Asians (32). While we would not expect perfect replication even if associations were tagging single, common, causal variant, we would expect practically none if they were synthetic, both because the underlying rare, causal alleles would be less likely to be shared among populations and because the particular LD configuration that allows for their tagging in one population would likely break down in others (33, 34). A second is that simulation studies suggest that synthetic associations are expected to have much lower MAF than typically observed among associations in GWAS (31). Moreover, these simulations suggest that, because synthetic association should capture only a fraction of the variance contributed by the tagged loci, having many synthetic associations would imply there being much more heritable variance than is known to be present in the population. A third, and perhaps most direct line of evidence, is that, to the best of our

knowledge, none of the studies that pursued fine-mapping around GWS associations have uncovered such synthetic associations (33, 35, 36). These arguments, together with other lines of evidence (e.g., 31) suggest that in practice synthetic associations are likely to be rare.

Perhaps a more plausible alternative is for an association to primarily tag one common, causal variant, with which it is in high LD, but also to pick up the effects of one or a few rare, causal variants, which are more poorly tagged. Under this scenario, we might expect the estimated contribution to variance to slightly overestimate the contribution of the dominant causal variant. To the best of our knowledge, this scenario has not been well characterized, making it difficult to assess how common it is or whether the overestimation would be substantial.

In summary, given what we now know, our assumption about the distribution of estimated variances among associations reflecting the distribution among causal loci seems sensible.

## 7. Inference

In this section, we describe how we used our model to make inferences based on GWAS results for height and body mass index (BMI). As we note in the Discussion, these inferences are meant as an illustration and do not incorporate the effects of demography and a few other factors (e.g., genotyping and errors in the estimation of effect sizes (24, 26)), which lie beyond the scope of this study.

### 7.1. The composite likelihood

Our inferences are based on a composite-likelihood approach. We begin by describing the composite-likelihood function and its maximization, when the loci detected by GWAS are strongly selected and can be described by the high-pleiotropy limit. In this case, we have shown that the distribution of variances among loci is insensitive to the distribution of selection coefficients, depends on a single parameter $v_s$, and is well approximated by the probability density

$$\rho(v) = \frac{2\exp\left(-2\sqrt{v/v_s}\right)}{v} \qquad \text{(A2.88)}$$

(Section 3.2). Further approximating the power in GWAS as a step function (see Section 6), we find that the probability density of sites that exceed a threshold $v^*$ can be approximated by

$$\mathrm{f}(v|v_s, v^*) = \frac{\rho(v)}{\int_{v>v^*}\rho(v)} = \frac{\exp(-2\sqrt{v/v_s})}{2v\ \mathrm{I}(2\sqrt{v^*/v_s})}, \qquad \text{(A2.89)}$$

where $\mathrm{I}(x) \equiv \int_{t>x}\exp(-t)/t$ (see Eq. A2.37). We therefore approximate the log-composite-likelihood of $v_s$ given the contributions to variance of the $K$ loci detected in a GWAS, $\{v_i\}_{i=1}^K$, by

$$\mathrm{LCL}(v_s|\{v_i\}_{i=1}^K, v^*) = \sum_{i=1}^{K}\log\left(\mathrm{f}(v_i|v_s)\right) =$$

$$= -\left(2/\sqrt{v_s}\right)\sum_{i=1}^{K}\sqrt{v_i} - K\log\left(\mathrm{I}\left(2\sqrt{v^*/v_s}\right)\right) - \sum_{i=1}^{K}\log\left(\sqrt{v_i}\right). \quad \text{(A2.90)}$$

It follows that the composite-likelihood is maximized when

$$\hat{v}_s = \mathrm{argmin}_{v_s}\left\{2\overline{\sqrt{v}}/\sqrt{v_s} + \log\left(\mathrm{I}\left(2\sqrt{v^*/v_s}\right)\right)\right\}, \quad \text{(A2.91)}$$

where $\overline{\sqrt{v}} \equiv \frac{1}{K}\sum_{i=1}^{K}\sqrt{v_i}$.

We also consider the models without pleiotropy and in which the degree of pleiotropy is a parameter. In the case without pleiotropy,

$$\rho(v) = \frac{2\exp(-2v/v_s)}{v} \quad \text{(A2.92)}$$

(see Section 3.2). By following the same steps, we find that the composite-likelihood is then maximized when

$$\hat{v}_s = \mathrm{argmin}_{v_s}\left\{2\bar{v}/v_s + \log\left(\mathrm{I}(2\,v^*/v_s)\right)\right\}, \quad \text{(A2.93)}$$

where $\bar{v} \equiv \frac{1}{K}\sum_{i=1}^{K} v_i$ . When the degree of pleiotropy $n$ is a parameter of the model, we find that

$$\rho_n(v) = \int_{a_1} \frac{2}{v} exp\left(-\frac{2v/v_s}{a_1^2/(a^2/n)}\right)\varphi_n(a_1|a) \quad \text{(A2.94)}$$

(see Section 3.2). Again, following the same steps, we find that the probability density of sites that exceed a threshold $v^*$ is

$$\mathrm{f}_n(v|v_s) = \frac{\rho_n(v)}{\int_{v>v^*}\rho_n(v)} \quad \text{(A2.95)}$$

and the log-composite-likelihood is

$$\mathrm{LCL}(v_s, n|\{v_i\}_{i=1}^{K}, v^*) = \sum_{i=1}^{K}\log\left(\rho_n(v_i)\right) - K\log\left(\int_{v>v^*}\rho_n(v)\right). \quad \text{(A2.96)}$$

In the latter case, we used numerical maximization to show that the composite-likelihood estimates for height and BMI converge to the high pleiotropy limit. Specifically, we maximized the composite-likelihood specifying an interval of $[1,1000]$ for *n*, where for both traits the estimates converged to the upper limit of 1000. While numerical optimization does not allow us

to specify an infinite interval, the likelihood function and maximal value for $n$=1000 are indistinguishable from those in the high-pleiotropy limit.

### 7.2. Determining $v^*$ and removing outliers

Our likelihood maximization requires us to specify the value of the threshold $v^*$. We choose this threshold based on the empirical distributions of the contributions to variance among genome-wide significant associations (Fig. A2.13a & b). Specifically, when the contributions to variance approach the lower boundary for discovery, we observe a decline in the density of loci. This is likely due to a gradual reduction in power and suggests that our approximation for power (as a step function) breaks down for these values of variance. We therefore choose thresholds that appear to be above this decline ($v^* = 1.4 \cdot 10^{-4} V_P$ for height and $v^* = 1.35 \cdot 10^{-4} V_P$ for BMI; Fig. A2.13a & b), resulting in the removal of 53 loci for height and 11 for BMI. We also examine how our estimates of $v_s$ depend on the choice of $v^*$, and find that they are much more sensitive to reducing the threshold than to increasing it; in fact, the estimates we obtain by increasing the threshold are within the confidence intervals of the estimate with the chosen thresholds (Fig. A2.13c & d). This analysis further supports our choice to exclude the loci with the lowest contribution to variance. For BMI, we also dropped the locus with the largest contribution to variance (near FTO), which appears to be an outlier (Fig. A2.13b) and has been suggested to be under balancing selection (37).

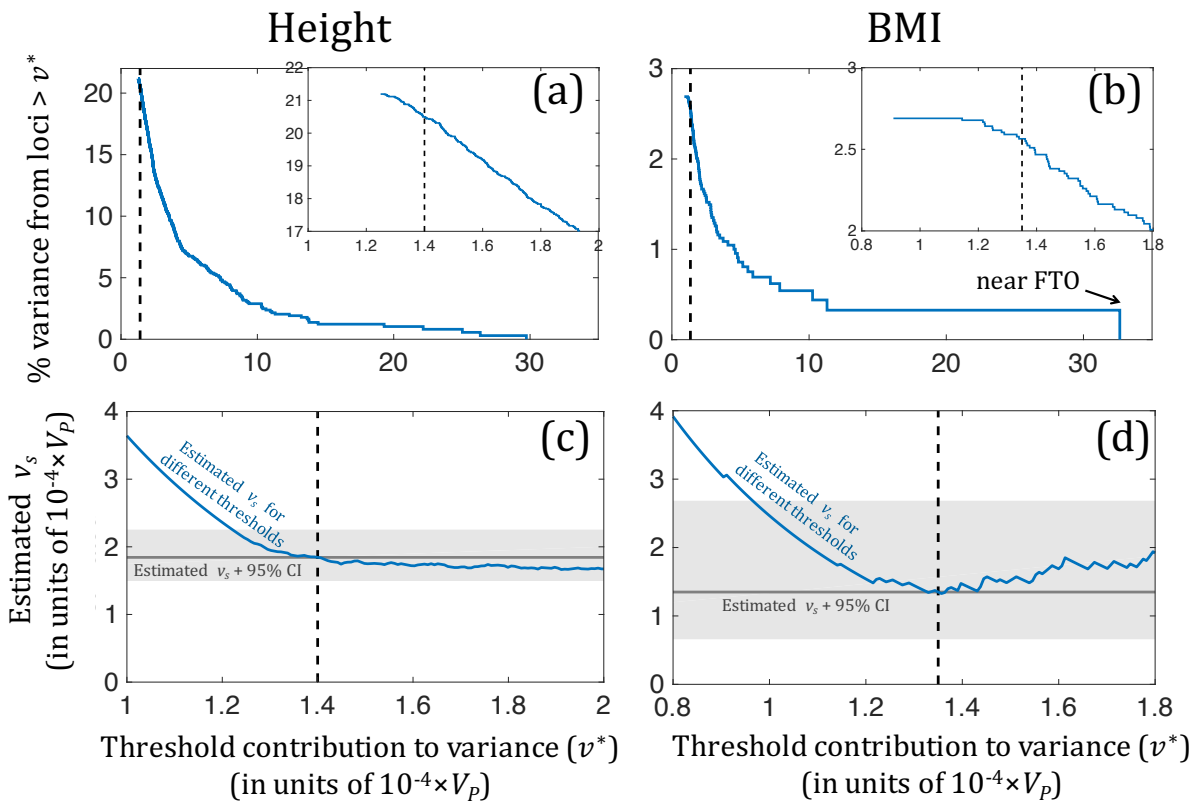



**Figure A2.13.** Determining $v^*$ and removing outliers. The total variance from significant associations as a function of the threshold contribution to variance, for height (a) and BMI (b). The insets show a close up of the lower range of contributions to variance, highlighting the decline in the density of discovered loci. Our chosen thresholds are shown by the dashed vertical line (in all graphs). Our estimates of $v_s$ as a function of the chosen threshold, for height (c) and BMI (d). When we increase the threshold, the estimates remain within the 95% CI of the estimate with our chosen threshold.

### 7.3. Estimating target size and explained variance

We estimate the target size and the variance explained, both for varying study size and total, based on our estimates of $v_s$. The population-scaled mutational input per generation from strongly selected loci, $2NU_s$, is estimated by

$$\widehat{2NU_s} = K / \int_{v>v^*} \rho(v|\hat{v}_s), \quad \text{(A2.97)}$$

(see Eq. A2.38) and the corresponding estimate for the target size is

$$\hat{L}_s = \widehat{2NU_s}/\widehat{2Nu}, \quad \text{(A2.98)}$$

where the estimate for the population scaled mutation rate per site per generation $\widehat{2Nu} \approx 0.5 \cdot 10^{-3}$ is based on heterozygosity (29). The explained variance corresponding to GWAS with study size $m$ is estimated by

$$\hat{\sigma}_s^2(m) = \widehat{2NU_s} \int_{v>v^*(m)} v\rho(v|\hat{v}_s) = K\left(\int_{v>v^*(m)} v\rho(v|\hat{v}_s) / \int_{v>v^*(m_0)} \rho(v|\hat{v}_s)\right), \quad \text{(A2.99)}$$

where we approximate the threshold corresponding to study size $m$ based on the study size, $m_0$, and threshold, $v^*$, in current GWAS, by

$$v^*(m) = v^* \cdot (m_0/m). \quad \text{(A2.100)}$$

To estimate the total variance arising from strongly selected loci, we simply set the threshold in Eq. A2.99 to 0.

### 7.4. Estimating confidence intervals

We use a combination of non-parametric and parametric bootstrap to estimate confidence intervals (CI). We use non-parametric bootstrap to estimate the CI for the model parameters $v_s$ and $L_s$: specifically, we perform 10,000 iterations, in which we resample the loci identified by GWAS and repeat the estimation of $v_s$. We use parametric bootstrap to estimate the confidence intervals in Fig. 5a, describing the explained variance as a function of threshold based on our

model. To that end, we rely on our model with the point estimates for $v_s$ and $L_S$, to generate 10,000 samples from GWAS with the specified threshold, and then calculate the total variance explained by these samples. We use a combination of non-parametric and parametric bootstrap to calculate the CI for model predictions, including the total variance, $\sigma_s^2$, and the explained variance, $\sigma_s^2(m)$, and number of loci as a function of study size (Fig. 5b & c). In this case, we generate 10,000 samples by: i) estimating $v_s$ based on a resampled set of GWAS loci (similar to the non-parametric procedure), and ii) using the estimated $v_s$ and corresponding $L_s$ to generate a GWAS hits above $v^*$ based on our model (similar to the parametric procedure); we then calculate the appropriate summary based on the latter samples. This two stage procedure is intended to capture the uncertainty generated by both the errors in estimating our basic model parameters and the noise generated by the stochastic processes underlying the number and variance at segregating loci that are yet to be discovered. The resulting estimates and CI are summarized in Table A2.2.

| Parameter | | Height | BMI |
|---|---|---|---|
| Contribution to variance per strongly selected locus, in units of the total phenotypic variance | $\hat{v}_s/V_P$ | 1.8 [1.5, 2.3]×$10^{-4}$ | 1 [0.6, 1.7] ×$10^{-4}$ |
| Expected study size required to capture 50% of the strongly selected variance | $m_{50\%}$ $(\approx 43V_P/\hat{v}_s)$ | 230 [190, 290] K | 420 [250, 770] K |
| Number of newly arising strongly selected mutations per generation in the population | $2NU_s$ | 2300 [1800, 3000] | 600 [300, 1900] |
| Mutational target size for strongly selected mutations | $L_s$ | 4.6 [3.6, 6.0] Mbp | 1.3 [0.6,3.8] Mbp |
| % contribution to phenotypic variance from strongly selected loci | $\sigma_s^2/V_P$ | 42 [39, 45] % | 7 [5, 10] % |
| Proportion of heritability from strongly selected loci | $\sigma_s^2/V_G$ $(= \sigma_s^2/h^2V_P)$ | 53 [49, 57] % | 13 [10, 21] % |

**Table A2.2.** Parameter estimates and their confidence intervals for height and BMI based on GWAS results; the heritability was assumed to be 0.8 for height and 0.5 for BMI (8, 10).

### 7.5. Testing goodness of fit

We use the Kolmogorov-Smirnov D statistic (38, 39) to test the goodness of fit of our models without pleiotropy and in the high pleiotropy limit. Since our parameter estimates are inferred from the data that we are testing against, we cannot rely on the standard tables for the p-values.

We therefore generate null distributions for the D statistic using parametric bootstrap based on our models. Specifically: i) we generate $10^5$ samples of $K$ significant loci based on the model under consideration, with the corresponding estimate of $v_s$, ii) we infer $v_s$ based on each sample, and iii) calculate the Kolmogorov-Smirnov D statistic between the distribution of variances for the $K$ loci in each sample and the corresponding theoretical distribution based on the $v_s$ inferred from that sample. The resulting distribution of D statistics corresponds to our null hypothesis, i.e., that the loci detected in GWAS arose according to our model, and specifically to the way we calculate the D statistic between the observed distribution of variances for the K detected loci and the theoretical distribution that we inferred based on these observations. We then calculate the D statistic, $D_r$, based on the real data and corresponding theoretical distribution, and estimate the one-sided p-value by

$$\hat{p}_{K-S} = \frac{\#\text{ simulated datasets with } D>D_r}{\#\text{ simulated datasets}}. \quad \text{(A2.101)}$$

Note that unlike the common case, here the inability to reject the null indicates that the data is consistent with our model.

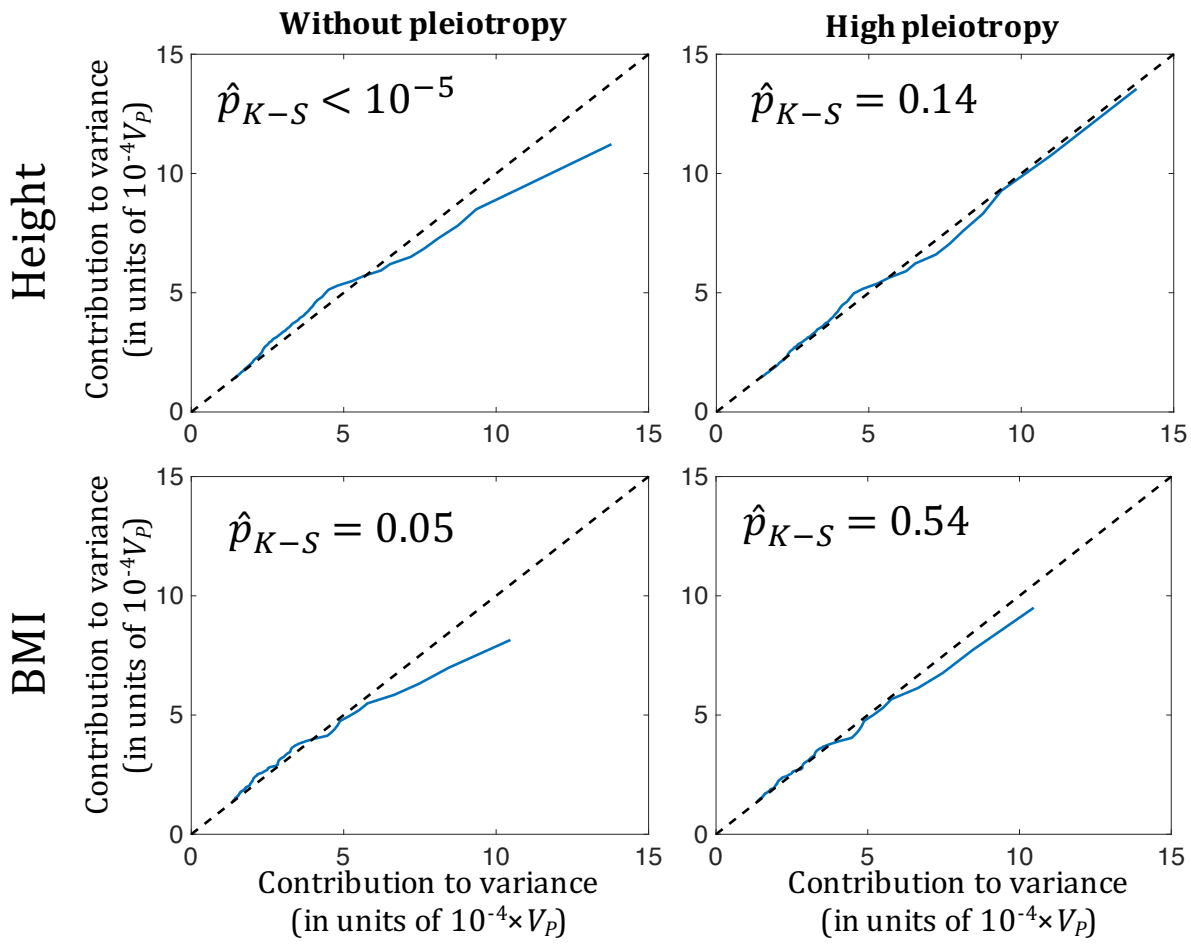



**Figure A2.14.** Q-Q plots comparing the distribution of variances among significant loci taken from the GWAS of height (10) and BMI (8) with the theoretical distributions inferred from these data, based on the models without pleiotropy (a) and in the high pleiotropy limit (b). These plots show that the model assuming high pleiotropy cannot be rejected for either trait and fits these data much better than the model without pleiotropy.

## 8. Consistency with other datasets and analyses

Here, we show that the results of our inference for height are consistent with findings of a recent GWAS based on exome genotyping; that our inferences for height and BMI are consistent with estimates of the heritability tagged by SNPs with MAF $> 1\%$ in the GWAS we used; and that our model is consistent with estimates about the relationship between effect size and MAF in these and other GWAS.

### 8.1. Exome association study of height

Marouli et al. (40) present an association study for height that was specifically designed to capture rare, exonic variants. They rely on the ExomeChip genotyping array (41), which includes the vast majority of protein-altering variants with MAF>0.1%, allowing them to directly (i.e., without imputation) test for associations among rare variants. Using a study size of more than 300,000 European individuals, they find over 400 genome-wide significant associations. Here we examine whether their findings are consistent with our inference based on the Wood et al. genome-wide, genotyping based GWAS for height (10).

In addition to protein altering variants, the ExomeChip includes some synonymous SNPs and ancestry informative markers, as well as all of the genome-wide significant associations listed in NHGRI from 2011. To avoid ascertainment biases, we consider only protein-altering variants, including non-synonymous, splice region, splice acceptor and stop codon variants. This leaves us with 250 of the Marouli et al. genome-wide significant associations. In addition, we apply the procedure described in Section 7.2, resulting in the removal of associations with contributions to variance below $v_E^* = 1.15 \cdot 10^{-4}\, V_P$, for which power is substantially diminished (Fig. A2.15a) ; this step leaves us with 147 associations.

Next, we compare the distribution of variances among the remaining 147 associations with our theoretical prediction, with the $v_S$ inferred from the Wood et al. data (Table A2.2) above the threshold $v_E^*$ (Fig. A2.15b). We do not consider the fit to the number of associations, because it depends on the mutational target size for protein-altering variants affecting height, which is unknown.

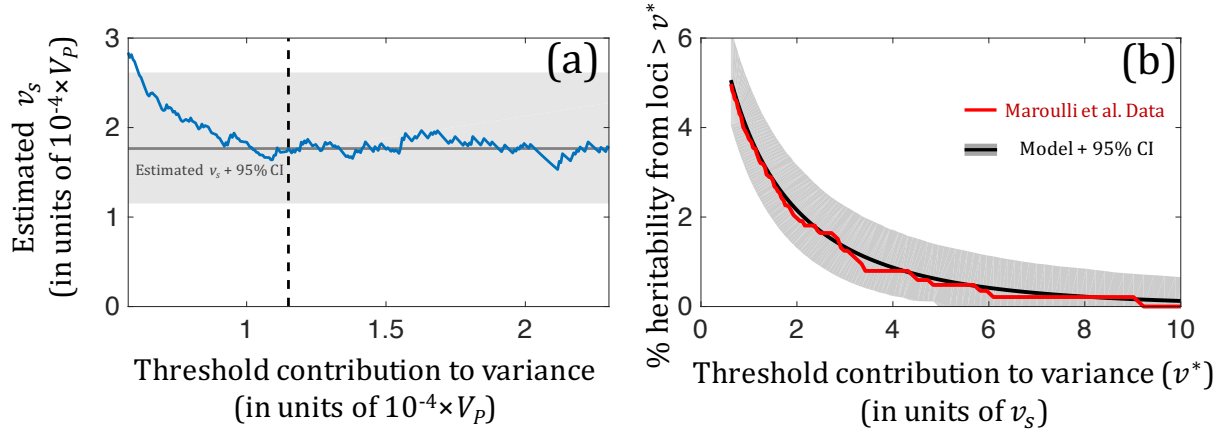



**Figure A2.15.** Comparing our inferences for height with the results of the Marouli et al. GWAS. (a) Choosing the threshold contribution to variance, $v_E^*$, above which our approximation for power applies; see Section 7.2 for details. (b) Comparing the predicted and observed distribution of variances above the threshold $v_E^*$. 95% CIs for our predictions are based on bootstrap; see Section 7.4 for details.

To test whether the observed distribution is consistent with our prediction, we calculate the Kolmogorov-Smirnov *D* statistic (38, 39) for this comparison, $D_r$, and ask whether we can reject our prediction based on the value of $D_r$. In approximating the null distribution of the *D* statistic, we must consider that: i) Some of the Marouli et al. associations might have been tagged by the genome-wide significant associations in Wood et al., which we relied upon in estimating $v_S$; this would lead to smaller values of the *D* statistic than if the two sets of associations were independent. ii) Our estimate of $v_S$ includes some statistical error, due to the finite set of

associations on which it relies. To account for these factors, we employ a parametric bootstrap procedure that mimics how the value of the *D* statistic arises, under the conservative scenario in which any of the associations from Marouli et al. could have been included in the data that we used in our inference. Specifically, we assume that the distribution of variances among loci follows the theoretical prediction with our estimate of $v_S$, and i) We sample 147 associations from the predicted distribution with threshold $v_E^*$, corresponding to the Marouli et al. associations. ii) Given the number, *k*, of these associations that fall above the threshold of the Wood et al. GWAS, $v_G^* = 1.4 \cdot 10^{-4}\, V_P$ (Section 7.2), we sample an additional $644 - k$ variants from the predicted distribution with threshold $v_G^*$. The resulting 644 simulated associations that fall above $v_G^*$ correspond to the Wood et al. associations. iii) We infer $\hat{v}_s$ based on these 644 variants, thus mimicking our inference procedure, and calculate the *D* statistic for our predicted distribution with $\hat{v}_s$ and the distribution based on the 147 simulated variants. iv) We repeat this procedure $10^5$ times to approximate the distribution of *D* statistic under our null, and estimate the one-sided p-value by

$$\hat{p}_{K-S} = \frac{\#\text{ simulated datasets with } D>D_r}{\#\text{ simulated datasets}}. \qquad \text{(A2.102)}$$

Doing so, we find that $\hat{p}_{K-S} = 0.99$, and thus, we cannot reject our predictions based on the data from Marouli et al. (40). This result indicates a good fit to their findings.

### 8.2. The heritability arising from common SNPs

Yang et al. (42, 43) estimate the heritability that is tagged by common SNPs (MAF>1%) in GWAS of several traits, including height and BMI. Here we ask whether their estimates are consistent with our inferences based on genome-wide significant (GWS) associations from the same GWAS.

First, we consider our inferences predicated on equilibrium demography. On this assumption, we predict that GWS associations would be under intermediate or strong selection, roughly corresponding to $S > 5$. Our estimates then suggest what proportion of variance arises from loci under this range of selection effects, where the rest of the variance is assumed to arise from loci under weaker selection. The proportion of variance that arises from sites with $S < 5$ and MAF $> 1\%$, $P_w(> 1\%)$, can be bound from above by the variance that would arise if they were all effectively neutral, $P_n(> 1\%)$. Further denoting the proportion of variance that arises from sites with $S > 5$ and MAF $> 1\%$ by $P_s(> 1\%)$, and the overall proportion of variance from sites with MAF $> 1\%$ by $P(> 1\%)$, we obtain the following requirement:

$$P_s(> 1\%) = P(> 1\%) - P_w(> 1\%) \geq P(> 1\%) - P_n(> 1\%). \quad \text{(A2.103)}$$

For height, Yang et al. estimate that $P(> 1\%) = 0.59$ (42), and our estimate for $P_n(> 1\%) = 0.45$. As Fig. A2.16 shows, so long as most of the estimated variance with $S > 5$ (53%) arises from loci with $S < 135$, the requirement in Eq. A2.103 will be easily met. For BMI, Yang et al. estimate that $P(> 1\%) = 0.5$ (42), and our estimate for $P_n(> 1\%) = 0.83$. The lower bound in Eq. A2.103 is therefore negative, implying that requirement A2.103 is met regardless of the distribution of selection coefficients for $S > 5$.

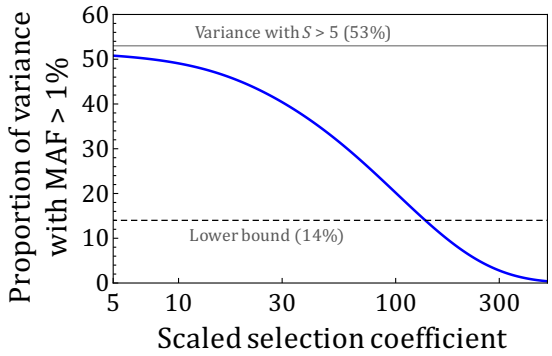



**Figure A2.16.** The Yang et al. (42, 43) estimate of the genetic variance in height arising from loci with $MAF > 1\%$ imposes weak constraints on the distribution of selection coefficients, assuming our estimate for the genetic variance with $S > 5$.

Next, we consider the results of our analysis in Section 9, incorporating the effects of recent changes in European population size. Our results suggest that GWS associations arise from loci with selection coefficients of $s \approx 10^{-3}$. We therefore ask whether the Yang et al. (42, 43) estimates are consistent with ours, when we attribute our equilibrium estimates of the proportion of variance arising from intermediate and strongly selected loci to selection coefficients of $s \approx 10^{-3}$, assuming that the remaining variance arises from loci under weaker or stronger selection (a more rigorous approach would be to account for demography in estimating the proportion of variance, but this extension lies beyond the scope of the current paper). The proportion of variance arising from sites under weaker selection with $\mathrm{MAF} > 1\%$ is bound from above by $P_n(> 1\%)$, whereas the corresponding proportion from sites under stronger selection can be vanishingly small. Denoting the proportion of variance arising from sites with $s \approx 10^{-3}$ and $\mathrm{MAF} > 1\%$ by $P_{10^{-3}}(> 1\%)$, we therefore obtain the following condition:

$$P(> 1\%) \geq P_{10^{-3}}(> 1\%) \geq P(> 1\%) - P_n(> 1\%). \quad \text{(A2.104)}$$

If we assume the Yang et al. (42) estimates for $P(> 1\%)$ and our estimates for $P_{10^{-3}}(> 1\%)$, Table A2.3 shows that this requirement is easily met for both height and BMI. More generally, our analysis illustrates that heritability estimates of this kind impose rather weak constraints on our inferences.

| | $P(>1\%)$ | | $P_{10^{-3}}(>1\%)$ | | $P(>1\%)-P_n(>1\%)$ |
|---|---|---|---|---|---|
| Height | 0.59 | ≥ | 0.48 | ≥ | 0.59−0.38=0.21 |
| BMI | 0.5 | ≥ | 0.12 | ≥ | 0.5−0.83=−0.33 |

**Table A2.3.** Consistency between the Yang et al. (42) estimates of the total variance arising from loci with MAF $> 1\%$ and our estimates of the variance arising from sites with $s \approx 10^{-3}$ and MAF $> 1\%$.

### 8.3. The relationship between SNP heterozygosity and effect size

More recent studies of the heritability tagged by SNPs in GWAS also make inferences about the relationship between effect sizes and MAF (44-46). Specifically, they assume that the relationship between the contribution of a site to variance, $v = 2a_1^2 x(1-x)$, and its MAF, $x$, takes the form

$$\mathrm{E}(v|x) \propto \big(x(1-x)\big)^{\alpha+1}, \tag{A2.105}$$

or equivalently, that

$$\mathrm{E}(a_1^2|x) \propto \big(x(1-x)\big)^{\alpha}, \tag{A2.106}$$

and they estimate the value of $\alpha$ from the data.

Provided a distribution of selection coefficients, $\mathrm{f}(S)$, Eq. A2.20 implies that in our model

$$\mathrm{E}(a_1^2|x) = \frac{\int_{a_1} a_1^2 \rho(x,a_1)}{\int_{a_1} \rho(x,a_1)} = \frac{\int_S \int_{a_1} a_1^2 \mathrm{f}(S)\tau(x|S)\eta(a_1|S)}{\int_S \int_{a_1} \mathrm{f}(S)\tau(x|S)\eta(a_1|S)}$$

$$= \frac{2w^2}{nN} \cdot \frac{\int_S S\,\mathrm{f}(S)\tau(\mathrm{x}|\mathrm{S})}{\int_S \mathrm{f}(S)\tau(\mathrm{x}|\mathrm{S})} = \frac{2w^2}{nN} \cdot \mathrm{E}(S|x). \tag{A2.107}$$

Thus, in our model, assuming the relationship of Eq. A2.105 (or A2.106) would imply that

$$\mathrm{E}(S|x) \propto \big(x(1-x)\big)^{\alpha} \tag{A2.108}$$

(See (45) for a similar derivation).

The aforementioned studies assume the relationship in Eq. A2.105 (or A2.106), without providing any evidence that this somewhat arbitrary functional form fits the data better than others, and show that values of $\alpha$ between -1 and 0 provide the best fit to data from GWAS of a variety of traits. To show that our model is in agreement with theirs, all we therefore need to do is to find distributions of selection coefficients, f($S$), that approximate the relationship of Eq. A2.108 for values of $\alpha$ between -1 and 0. In Fig. A2.17, we assume that selection coefficients follow a Gamma distribution, where we vary its expectation and variance. As expected, $\mathrm{E}(S|x)$ monotonically decreases as $x$ increases. When $\mathrm{E}(S) \ll 1$ or the coefficient of variation $\mathrm{C_V^2}(S) \ll 1$, $\mathrm{E}(S|x)$ varies minimally with $x$ and can approximated by Eq. A2.108 with $\alpha = 0$. In other cases, $E(S|x)$ varies more substantially with $x$. When we approximate those cases using Eq. A2.108, we obtain a range of $\alpha$ values between $-1$ and 0. Thus, our model appears to be consistent with the values of $\alpha$ reported in (44-46). Our inferences for height and BMI are not very informative about the distribution of selection coefficients and are therefore not comparable with estimates of $\alpha$.

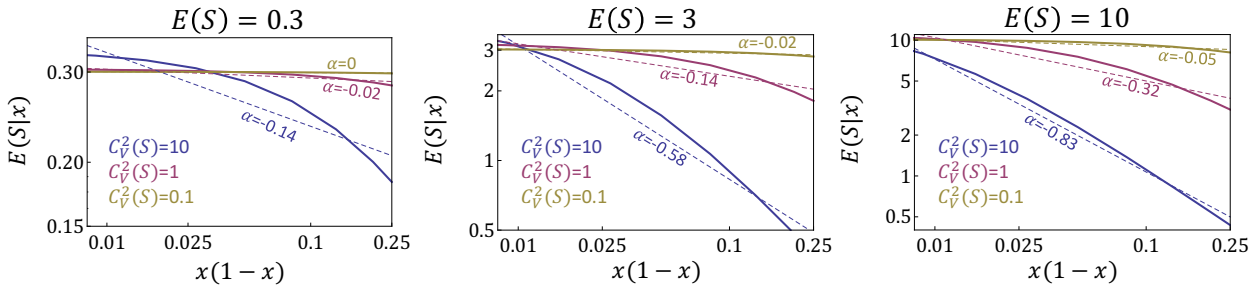



**Figure A2.17.** The relationship between effect size, or equivalently, selection coefficient, and MAF, shown on a log-log scale. Selection coefficients are gamma-distributed, with $\mathrm{E}(S) = 0.3, 3, 10$ and shape parameters $k = 0.1, 1, 10$. $\mathrm{E}(S|x)$ was approximated using the functional form $\mathrm{E}(S|x) \propto \big(x(1-x)\big)^{\alpha}$ (Eq. A2.108), by taking the values of $\log\big(\mathrm{E}(S|x)\big)$ and $\log\big(x(1-x)\big)$ on a grid of $x$ values, $x = 0.5 \cdot 10^{-i/4}$ with $i = -8, -7, \ldots, 0$, and preforming least-square linear regression.

## 9. The effects of demographic history

While our theoretical results were derived on the assumption of a panmictic population of constant size, the evolutionary history of human populations sharply deviates from these simplifying assumptions. Notably, most large GWAS, including the studies of height (10) and BMI (8) that we use to test our predictions, have been performed in predominantly European populations, which are known to have experienced dramatic changes in their effective population size, including an Out-of-Africa bottleneck about ~100 KYA and explosive population growth over the past ~5 KY (47-50). These changes in population size have dramatically impacted the frequencies of neutral and selected alleles (47-49, 51-53), and are therefore expected to have had a substantial impact on the architecture of quantitative traits (52, 53). These considerations raise several questions about the interpretation of the fit between our predictions and GWAS data. Notably, how will these historical changes in population size affect our prediction, and specifically, why do our equilibrium predictions fit GWAS data despite the dramatic historical changes in population size? While a comprehensive treatment of these questions warrants a study in itself, we briefly address them here.

Even with changing population size, our results for the dynamics at segregating sites should still hold. Notably, we would expect the mean phenotype in the population to maintain the optimal phenotypic value, because any displacement from the optimum would be quickly adjusted by small changes to allele frequencies at numerous loci (see Section 4.4). As a result, the dynamics at individual sites would be decoupled, and well approximated by the first two moments of change in allele frequency described in Eqs. 5 and 6, which appear in the Results section of Chapter 2. In particular, the first moment would correspond to under-dominant selection, and the selection coefficient would be proportional to the size of the allele in the $n$-dimensional trait

space (as described in Eq. 7, which appears in the Results section of Chapter 2). We can therefore study the effect of historical changes in population size on allele frequencies with simulations, using a fixed (not population-scaled) selection coefficient with under-dominance, and having the population size change over time.

To this end, we modify the simulation from Simons et al. (53) to incorporate under-dominance, and the historical changes in the effective population size of European populations inferred by Schiffels and Durbin (50) (Fig. A2.18). In brief, we simulate a bi-allelic site in a diploid, panmictic population, in which mutations, with selection coefficient *s*, arise at rate $u = 1.25 \cdot 10^{-8}$ per bp per generation (5, 50), and the next generation derives from Wright-Fisher sampling and fecundity selection. The simulation begins 150K generations ago (corresponding to 4.5 MYA with a generation time of 30 Y, as assumed by (50)), with a burn-in period with a constant population size of 14,448. In accordance with the Schiffels and Durbin inferences (50), changes in population size begin 55,940 generations ago (corresponding to 1.7 MYA). Specifically, we piece together the MSMC inferences from two and four haplotypes of European individuals (CEU) from HapMap project (54), where the four haplotype MSMC captures the bottleneck and recent growth and is used for times <170 KYA, and the two haplotype MSMC captures more ancient times and is used for times >170 kya (see Fig. A2.18). The derived allele frequency is recorded at the last generation corresponding to the present. The software and documentation can be found at https://github.com/sellalab/GenArchitecture.

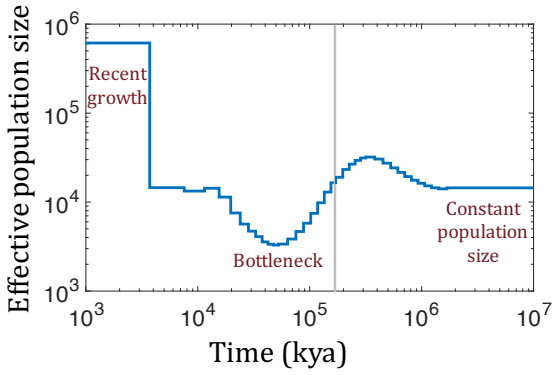



**Figure A2.18.** Changes in population size in the history of Europeans, as inferred by Schiffels and Durbin using MSMC (50). The cutoff between the two and four haplotype MSMC inferences is marked by the gray line.

We rely on such simulations to study how changes in populations size will affect the genetic architecture of a trait under the assumptions of our model. To this end, we consider a grid of selection coefficients: $s = 10^{-i/8}$, $i = 8, 9 \ldots, 40$, where for each selection coefficient, we run $15 \cdot 10^6$ simulations. In this way, we obtain numerical approximations for the expected site frequency spectrum corresponding to each selection coefficient, which replaces the term $2Nu \cdot \tau(x|S)$ in our expressions for summaries of genetic architecture (Section 3). We further assume the high pleiotropy limit form for the distribution of effect sizes on the focal trait corresponding to a given selection coefficient (i.e., Eq. 11, which appears in the Results section of Chapter 2).

We first consider how demography affects the distribution of genetic variances among sites with different selection coefficients (Fig. A2.19a). The expected contribution per site (including both sites that are segregating and monomorphic) peaks around a selection coefficient of $s \approx 10^{-3}$ and, as in the case with constant population size (Fig. 2a), when the strength of selection increases, it appears to approach a plateau (Fig. A2.19a). The distribution of variances among sites, however, is dramatically affected by changes in population size: for selection coefficients around $s \approx 10^{-3}$, a much greater proportion of variance comes from sites with large contributions than from those with both weaker and stronger selection coefficients (Fig. A2.19b).

This behavior contrasts with the case of a constant population size, where for sufficiently strong selection ($S > 5$), the distribution of variances among sites is insensitive to the strength of selection (see Fig. 3b).

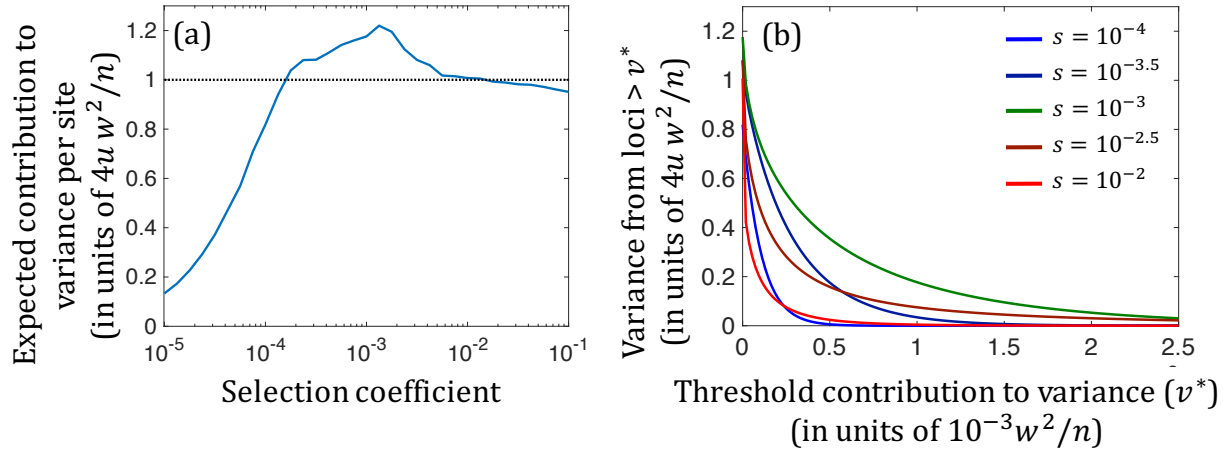



**Figure A2.19.** The joint effects of selection and changes in populations size (as inferred for Europeans by Schiffels and Durbin (50)) on the distribution of genetic variances among sites. (a) The expected contribution to variance per site, both segregating and monomorphic, as a function of the (unscaled) selection coefficient. Variance is measured in units of $4u\,w^2/n$, the equilibrium expectation for a strongly selected site. (b) The cumulative variance arising from sites with contributions above a threshold (y-axis) as a function of the threshold (x-axis); cumulative variance is measured in units of $4u\,w^2/n$, while the threshold in units of $10^{-3}w^2/n$.

As we establish below, these findings can be understood as follows. The segregating sites with the largest contribution to current genetic variance are due to mutations with $s \approx 10^{-3}$ that arose shortly before or during the Out-of-Africa bottleneck. Such mutations were under strong selection (i.e., with $2N_e s \approx 50$) before the bottleneck, but with the drop to an effective population size of $N_e \approx 4000$ during the bottleneck, they experienced more relaxed selection (with $2N_e s \approx 10$), allowing some of them to ascend to higher frequencies. The durations of subsequent increases in population size, and of explosive growth in particular, were too short to allow for a substantial reduction in their frequencies (e.g., a mutation with $s = 10^{-3}$ that reached 20% frequency by the end of the bottleneck, 15 Kya, would have an expected frequency of 18%

at present). As a result, these mutations would have large contributions to variance at present. Moreover, their site frequency spectrum and distribution of contributions to variance are well approximated by assuming a population size of $N_e \approx 5000$ – roughly the geometric mean of populations sizes from the beginning of the bottleneck to the present – and thus to scaled selection coefficients of $2N_e s \approx 10$.

Extant segregating mutations under substantially stronger selection are expected to be much younger. They therefore tend to have arisen after the bottleneck, when the population size was considerably larger. As a result, they have much lower frequencies and per segregating site contributions to variance at present. The larger population size, however, will also increase the mutational input and thus the number of extant segregating sites; so long as selection is sufficiently strong, these effects balance each other such that the per site contribution to variance, counting both segregating and monomorphic sites, remains insensitive to changes in population size (53). In turn, extant segregating mutations under substantially weaker selection are expected to contribute much less variance per site (considering either segregating sites alone or all sites) primarily because of their smaller effect sizes, which is the same reason that applied in the case with a constant population size (see Fig. 2a).

We find support for this verbal argument when we relate the results of our simulations with the findings from GWAS. To do so, we follow the same reasoning that we applied to the case with constant population size (see Discussion). Namely, based on the distribution of variances (Fig. A2.19a), we would expect sites with selection coefficients around $s \approx 10^{-3}$ to be the first to be discovered in GWAS. Further assuming that such sites account for most associations discovered in GWAS and that their distribution of variances corresponds to $N_e = 5000$, we can use our estimates of $v_s$ for height and BMI to calculate the parameter $w^2/n$ ($= ½N_e v_s$) for these

traits. This approach allows us to plot the putative distribution of variances among sites that exceed the study thresholds, $v^*$, for different selection coefficients (Fig. A2.20a & b). Doing so, we find that the observed and fitted distributions are well approximated by the distributions for sites with $s \approx 10^{-3}$, thus supporting our premise that most of the explained variance arises from such sites, and that their distribution of variances is well approximated by assuming a constant population size of $N_e \approx 5000$. Our simulations also suggest that the proportion of variance explained for sites with $s \approx 10^{-3}$ is much greater than the proportion for sites under weaker or stronger selection (Fig. A2.20c & d), and should therefore also be greater than the total proportion of variance explained by these GWAS. This expectation accords with our findings as well, with our simulations suggesting that the proportion of variance explained for sites with $s \approx 10^{-3}$ is $\sim$40% for height and $\sim$30% for BMI (Fig. A2.20c & d) compared to a total proportion of $\sim$25% for height and $\sim$5% for BMI in these GWAS (8, 10).

Examining the expected MAF and allelic ages at sites that we predict to have been identified by these GWAS lends further support to our interpretation (Fig. A2.20e-h). Notably, we find that the MAF for sites with $s \approx 10^{-3}$ that are predicted to have been identified by these studies are similar to those that are observed (Fig. A2.20e & f). Moreover, when we examine the ages of mutations at detected sites, we find that mutations at sites with $s \approx 10^{-3}$ are predicted to have originated during or shortly before the OoA bottleneck (Fig. A2.20g & h).

In summary, our analyses suggest that the bulk of associations identified in the GWAS for height and BMI tag segregating mutations with $s \approx 10^{-3}$, which originated shortly before or during the OoA bottleneck. As a result, we would expect the distribution of variances among these sites to be well approximated by our equilibrium predictions corresponding to an effective population of $N_e \approx 5000$. This finding provides an explanation for why our equilibrium predictions fit the

findings of GWAS in Europeans, despite our ignoring the dramatic changes in population size during their recent evolutionary past.

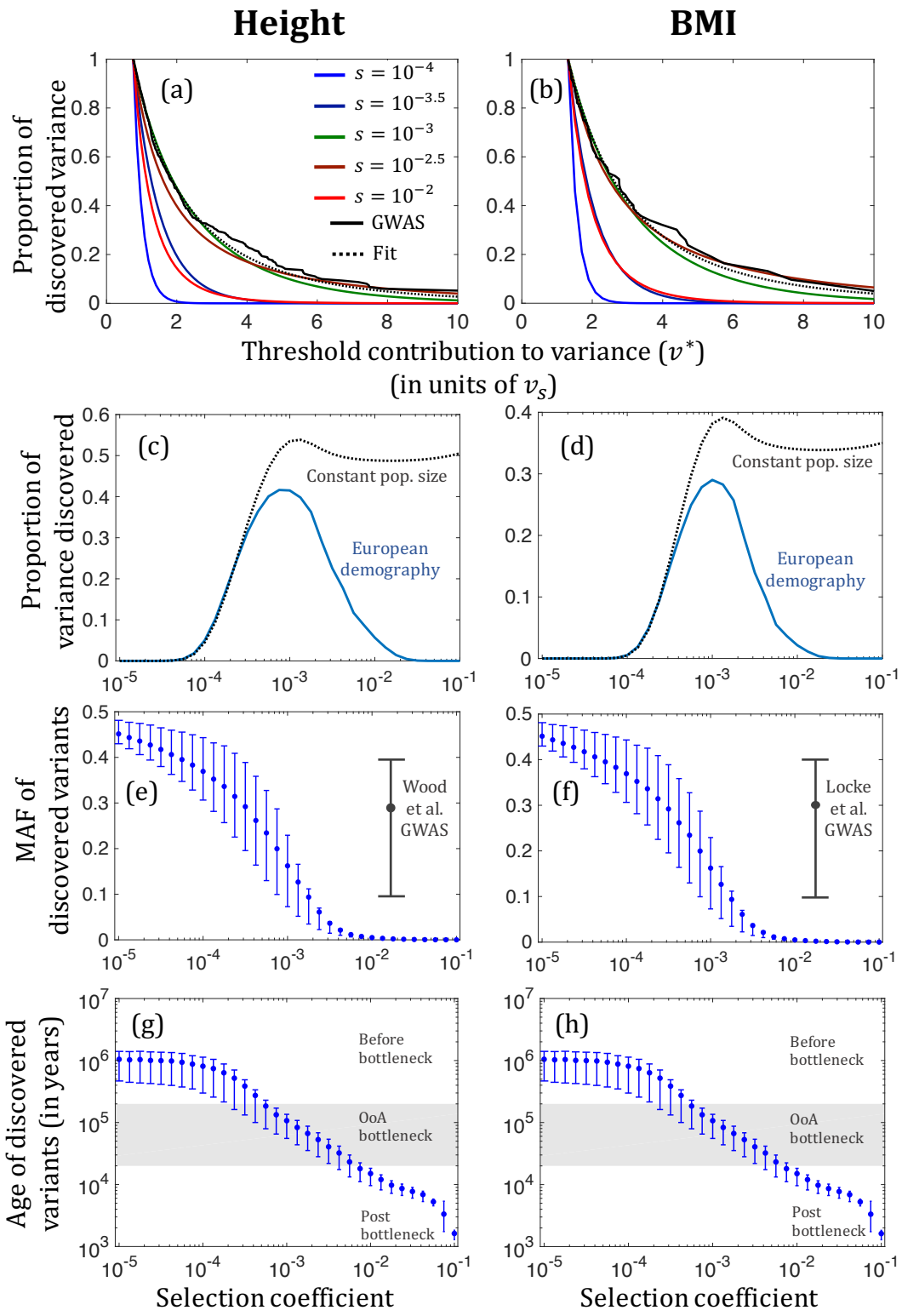


**Figure A2.20.** Comparison of the results of simulations with European demography with our inferences and the findings from GWAS for height and BMI (8, 10). (a & b) The distribution of variances among discovered loci. For each selection coefficient, the proportion of variance at the study threshold is set to 1. Simulation-based distributions are in color; the empirical distributions are in solid black; and our inferred fits are in dashed black. (c & d) The expected proportion of variance explained in GWAS as a function of the selection coefficient, based on simulations and on the equilibrium model with a constant population size of $N_e = 5{,}000$. (e & f) Comparison of MAF of discovered sites as a function of selection coefficient in simulations with the MAF observed for GWS associations in GWAS. (g & h) The age of mutations at discovered sites as a function of selection coefficient based on simulations. In (e–h), points correspond to the mean and whiskers span the 1st to 3rd quartiles of the distribution.

## 10. The effects of genotyping

Another implication of the demographic effects that we discussed in the last section (Section 9) pertains to the reliance on genotyping rather than resequencing in GWAS. As we reviewed in Section 6, current genotyping-based GWAS typically consider only loci with MAF $> 1\%$, for which imputation is currently quite accurate, at least in Europeans (24). Even if loci below that frequency were imputed with perfect accuracy, however, they would only be detected in a GWAS if they exceed the threshold contribution to variance for that study. Thus, loci at which the minor allele is rare would only be detected if they had very large effect sizes, which in our model implies very strong selection. For example, assuming a constant population size, if a re-sequencing study captured 25% of the heritable variance, a genotyping study with the same sample size would suffer a $\geq 50\%$ decrease in explained heritability only if $S \geq 200$ (Fig. A2.21a & b)For an effective population size of $2 \cdot 10^4$ for humans (50), that implies an enormous fitness cost of $s \geq 0.5\%$ (in heterozygotes) for the minor allele.

Our results for European demographic history suggest that only a small proportion of genetic variance can arise from loci that fall below the current MAF imputation threshold but have sufficiently large effect sizes to exceed the variance discovery thresholds of current GWAS. To illustrate that, we relied on our simulation results and estimates of $v_s$ for height and BMI, to calculate the proportion of variance arising from sites with MAF $< 1\%$ and contribution to variance $> v^*$. We find this proportion to be greater than 0 only for selection coefficients between $s = 0.3 \cdot 10^{-2}$ and $s = 2 \cdot 10^{-2}$, but even within this range, such loci account for less than ~6% of the expected variance for height and 2% of the variance for BMI (Fig. A2.21c & d). This suggests that the common reliance on genotyping in current GWAS for quantitative traits entails minimal loss in the discovery of associations relative to resequencing. Moreover, this is

likely to remain the case even when GWAS sizes substantially increase, as long as such increases are accompanied by reasonable increases in the size of imputation panels.

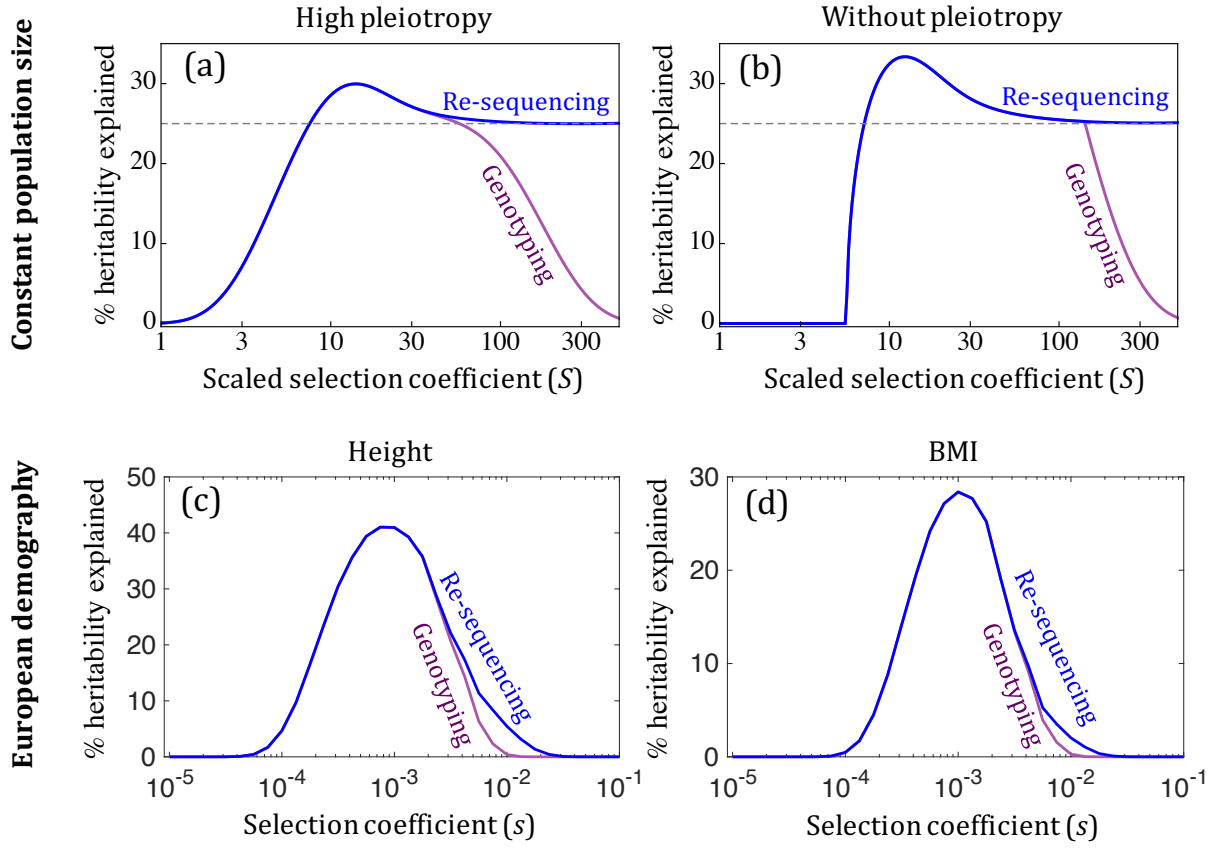



**Figure A2.21.** The heritability explained in resequencing and genotyping studies as a function of selection coefficient. (a-b) The case with a constant population size, in the high pleiotropy limit (a) and without pleiotropy (b). The study size was chosen such that a resequencing study would capture 25% of the strongly selected variance: implying a study size of $\sim 16 V_P/v_s$ in the highly pleiotropic limit (a), and a study size of $\sim 43 V_P/v_s$ without pleiotropy (b). (c-d) The case with European demographic history and high pleiotropy, using our estimates of $v_s$ for height (c) and BMI (d) (see Section 9 for details).

## 11. Glossary of notation

| | |
|---|---|
| $\vec{r}$ | n-dimensional phenotype |
| $\mathrm{W}(\vec{r})$ | Absolute fitness |
| $w$ | Scale of selection |
| $n$ | Number of traits (dimension) |
| $\vec{a}$ | A mutation's $n$-dimensional effect size |
| $a_1$ | A mutation's effect size on focal trait |
| $U$ | Haploid mutation rate per generation |
| $\sigma^2$ | Phenotypic variance in a trait |
| $K$ | Number of segregating sites |
| $\varphi_n(a_1 \mid a)$ | Distribution effect sizes on focal trait conditional on overall effect size |
| $S = \frac{N\mathrm{a}^2}{2\mathrm{w}^2}$ | Scaled selection coefficient |
| $\eta(a_1 \mid S)$ | Distribution of effect sizes on focal trait conditional on $S$ |
| $q$ | Derived allele frequency |
| $p$ | Ancestral allele frequency, $p = 1 - q$ |
| $\tau(q \mid S)$ | The sojourn time for a mutation with scaled selection coefficient $S$ |
| $v$ | Contribution to variance from a site $\left(v = 2a_1^2 q(1-q)\right)$ |

| | |
|---|---|
| $v_s$ | Expected contribution of a strongly selected site to variance ($v_s = 2w^2/nN$). |
| $\mathrm{E}(V\|S)$ | Expected contribution to genetic variance from sites with $S$ |
| $\mathrm{E}(K\|S)$ | Expected number of segregating sites with $S$ |
| $\rho(v)$ | Density of segregating sites contributing variance $v$ |
| $\mathrm{G}(v^*)$ | Proportion of variance from sites with contribution to variance > $v^*$ |
| $m$ | GWAS study size |
| H | Power to identify a locus in GWAS |

## 12. Additional figure

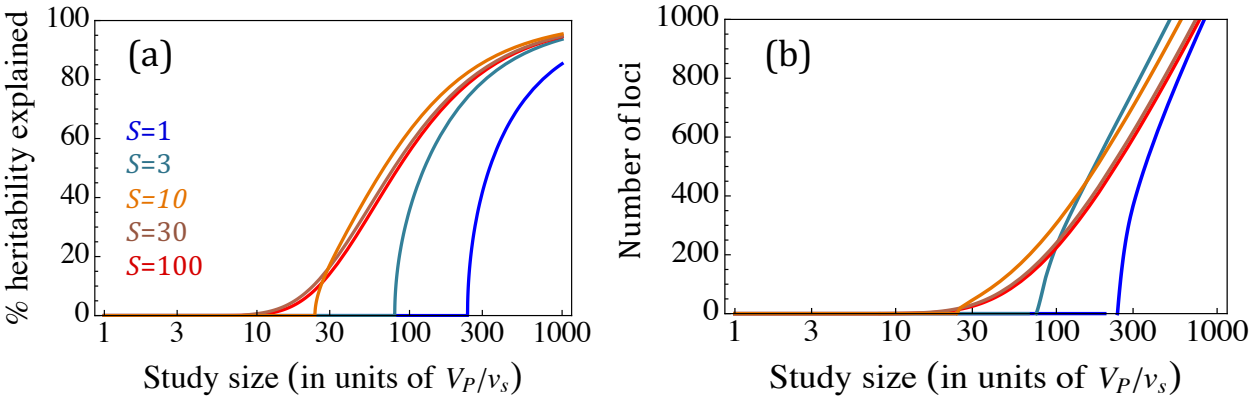



**Figure A2.22.** The proportion of heritability (a) and the number of variants per Mbp (b) identified in GWAS as a function of study size, in the case without pleiotropy ($n = 1$); see Section 3 for derivations. This figure is equivalent to Fig. 4 from Chapter 2, which describes the case with pleiotropy ($n \gg 1$).

## References


1. Lande R (1975) The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genet Res* 26(3):221-235.

2. Poon A & Otto SP (2000) Compensating for our load of mutations: Freezing the meltdown of small populations. *Evolution* 54(5):1467-1479.

3. Ewens WJ (2004) *Mathematical population genetics 1: I. Theoretical introduction* (Springer Science & Business Media).

4. Keightley PD & Hill WG (1988) Quantitative genetic variability maintained by mutation-stabilizing selection balance in finite populations. *Genet Res* 52(01):33-43.

5. Kong A*, et al.* (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412):471-475.

6. Ward LD & Kellis M (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337(6102):1675-1678.

7. Perry JRB*, et al.* (2014) Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* 514(7520):92-97.

8. Locke AE*, et al.* (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518:197-206.

9. Scott RA*, et al.* (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet* 44(9):991-1005.

10. Wood AR*, et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 46(11):1173–1186.

11. Schork NJ, Murray SS, Frazer KA, & Topol EJ (2009) Common vs. Rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19(3):212-219.

12. Hunt KA*, et al.* (2013) Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* 498(7453):232-235.

13. Do R*, et al.* (2015) Exome sequencing identifies rare ldlr and apoa5 alleles conferring risk for myocardial infarction. *Nature* 518(7537):102-106.

14. Purcell SM*, et al.* (2014) A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506(7487):185-190.

15. Chen JA, Peñagarikano O, Belgard TG, Swarup V, & Geschwind DH (2015) The emerging picture of autism spectrum disorder: Genetics and pathology. *Annu Rev Pathol Mech Dis* 10:111-144.

16. Uhlenbeck GE & Ornstein LS (1930) On the theory of the brownian motion. *Phys Rev* 36(5):823.

17. De Vladar HP & Barton N (2014) Stability and response of polygenic traits to stabilizing selection and mutation. *Genetics* 197(2):749-767.

18. Jain K & Stephan W (2017) Rapid adaptation of a polygenic trait after a sudden environmental shift. *Genetics* 207(3).

19. Lande R (1976) Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30(2):314-334.

20. Charlesworth B (2013) Stabilizing selection, purifying selection, and mutational bias in finite populations. *Genetics* 194(4):955-971.

21. Leffler EM*, et al.* (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339(6127):1578-1582.

22. Denny JC*, et al.* (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotech* 31(12):1102-1111.

23. Martin G & Lenormand T (2006) A general multivariate extension of fisher's geometrical model and the distribution of mutation fitness effects across species. *Evolution* 60(5):893-907.

24. Sham PC & Purcell SM (2014) Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 15(5):335-346.

25. Dickson SP, Wang K, Krantz I, Hakonarson H, & Goldstein DB (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol* 8(1):e1000294.

26. Evangelou E & Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* 14(6):379-389.

27. Visscher PM, Brown MA, Mccarthy MI, & Yang J (2012) Five years of GWAS discovery. *Am J Hum Genet* 90:7-24.

28. Pritchard JK & Przeworski M (2001) Linkage disequilibrium in humans: Models and data. *Am J Hum Genet* 69(1):1-14.

29. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526(7571):68-74.

30. Yang J*, et al.* (2012) Conditional and joint multiple-snp analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 44(4):369-375.

31. Wray NR, Purcell SM, & Visscher PM (2011) Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol* 9(1):e1000579.

32. He M*, et al.* (2015) Meta-analysis of genome-wide association studies of adult height in east asians identifies 17 novel loci. *Hum Mol Genet* 24(6):1791-1800.

33. Anderson CA, Soranzo N, Zeggini E, & Barrett JC (2011) Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol* 9(1):e1000580.

34. Liu JZ*, et al.* (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* 47(9):979-986.

35. Spain SL & Barrett JC (2015) Strategies for fine-mapping complex traits. *Hum Mol Genet* 24(R1):R111-R119.

36. Zheng J*, et al.* (2017) Haprap: A haplotype-based iterative method for statistical fine mapping using GWAS summary statistics. *Bioinformatics* 33(1):79-86.

37. Liu X*, et al.* (2015) Signatures of natural selection at the FTO (fat mass and obesity associated) locus in human populations. *PLoS One* 10(2):e0117093.

38. Genest C & Remillard B (2008) Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Ann Inst H Poincare Probab Statist* 44(6):1096-1127.

39. Stute W, Manteiga WG, & Quindimil MP (1993) Bootstrap based goodness-of-fit-tests. *Metrika* 40(1):243-256.

40. Marouli E*, et al.* (2017) Rare and low-frequency coding variants alter human adult height. *Nature* 542(7640):186-190.

41. Grove ML*, et al.* (2013) Best practices and joint calling of the humanexome beadchip: The charge consortium. *PLoS One* 8(7):e68095.

42. Yang J*, et al.* (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 47(10):1114-1120.

43. Yang J*, et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565-569.

44. Speed D*, et al.* (2017) Reevaluation of snp heritability in complex human traits. *Nat Genet* 49(7):986-992.

45. Schoech A*, et al.* (2017) Quantification of frequency-dependent genetic architectures and action of negative selection in 25 uk biobank traits. *bioRxiv*.

46. Zeng J*, et al.* (2017) Widespread signatures of negative selection in the genetic architecture of human complex traits. *bioRxiv*.

47. Wall JD & Przeworski M (2000) When did the human population size start increasing? *Genetics* 155(4):1865-1874.

48. Coventry A*, et al.* (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* 1:131.

49. Tennessen JA*, et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64-69.

50. Schiffels S & Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46(8):919-925.

51. Gazave E, Chang D, Clark AG, & Keinan A (2013) Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics* 195(3):969-978.

52. Lohmueller KE (2014) The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet* 10(5):e1004379.

53. Simons YB, Turchin MC, Pritchard JK, & Sella G (2014) The deleterious mutation load is insensitive to recent population history. *Nat Genet* 46(3):220-224.

54. The International Hapmap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311):52-58.

# Appendix 3

## Contents



## 1. The Likelihood

Our previous work (Chapter 2) on quantitative traits under pleiotropic, stabilizing selection showed that alleles are expected to experience underdominant selection. The expected change in allele frequency, x, is

$$E(\Delta x) = -sx(1-x)\left(\frac{1}{2}-x\right) \quad \text{(A3.1)}$$

The expected variance of the change in allele frequency is the standard drift term

$$V(\Delta x) = \frac{x(1-x)}{2N_e}, \quad \text{(A3.2)}$$

with $N_e$ being the effective population size. For a constant population size, we can use the diffusion approximation to calculate the sojourn time $\tau(x|s)$. Under the infinite sites model, the SFS can be approximated as

$$p(x|s) = 2N_e u \cdot \tau(x|s) \quad \text{(A3.3)}$$

with u being the mutation rate per site. $p(x|s)$ represents the probability density of a site with selection coefficient s to be at frequency x. Though $p(x|s)$ does not include it explicitly, there is a point mass of sites at $x = 0$, which are not segregating.

Our work also established the distribution of effect sizes conditional on the selection coefficient for traits under pleiotropic, stabilizing selection. In the limit of high pleiotropy, this distribution can be written as

$$p_c(a|s) = \frac{1}{\sqrt{2\pi cs}} \exp\left(-\frac{1}{2}\frac{a^2}{cs}\right) \quad \text{(A3.4)}$$

with c being a constant of proportion, converting selection to units of effect size.

$p(x|s)$ and $p_c(a|s)$ allow us to write an explicit likelihood function for the distribution of selection coefficients of newly arising mutations, $f(s)$ and c, as a function of the co-distribution of frequencies and effect sizes. We define the conditional distribution of alleles at frequency x and effect size a as

$$p_c(x, a|s) = p(x|s)p_c(a|s). \quad \text{(A3.5)}$$

We can then write the likelihood as

$$L(f(s), c|\{x, a\}) = \prod_i \frac{\int_s p_c(x,a|s)f(s)}{\int_s n_c(v^*,s)f(s)} \quad \text{(A3.6)}$$

with

$$n_c(v^*, s) = \int_{x,a|v>v^*} p_c(x, a|s) \quad \text{(A3.7)}$$

being probability that a site with selection coefficient s will be captured by a GWAS with a threshold contribution $v^*$. Where, like in Chapter 2, we assume GWAS captures all alleles with contributions to variance $v = 2a^2x(1 - x)$ above a threshold contribution $v^*$.

Only a limited band of selection coefficients can be captured in GWAS, because the effect sizes of small selection coefficients are too small and because large selection coefficients drive alleles to very low frequencies (see forward Appendix 3, Section 7). Therefore, the distribution of selection coefficients of newly arising mutations, $f(s)$, invariably includes information about ranges of selection coefficients for which GWAS give us no information. This lack of information would lead to singularities and artifacts in any attempt to infer $f(s)$.

We therefore prefer to recast the likelihood in terms of the distribution of selection coefficients at the GWAS hits themselves. The distribution of selection coefficients at GWAS hits will allow us to infer $f(s)$ in the relevant band of selection coefficients and is of intrinsic value by itself. For example, it can serve as a null model for tests of directional selection.

We define the distribution of selection coefficients at the GWAS hits as

$$g(s) = \frac{n_c(v^*,s)f(s)}{\int_s n_c(v^*s)f(s)} \tag{A3.8}$$

and we can therefore write a likelihood for it as

$$L(g(s), c|\{x, a\}) = \prod_i \frac{\int_s p_c(x,a|s)f(s)}{\int_s n_c(v^*,s)f(s)} = \prod_i \int_s \frac{p_c(x,a|s)}{n_c(v^*,s)} g(s)$$

$$= \prod_i \int_s P(x_i, a_i|v^*, s)g(s) \tag{A3.9}$$

with

$$P_c(x, a|v^*, s) \equiv \frac{p_c(x,a|s)}{n_c(v^*,s)} \tag{A3.10}$$

being the co-distribution of frequencies and effect sizes conditional on discovery in GWAS. Note that for an infinitely large GWAS, that is for $v^* = 0$, $P_c(x, a|v^*, s) = p_c(x, a|s)$ and $f(s) = g(s)$.

The log likelihood is therefore

$$LL(g(s), c|\{x, a\}) = \sum_i \log\left(\int_s P_c(x_i, a_i|v^*, s)g(s)\right). \quad \text{(A3.11)}$$

This form of the log likelihood is quite general and many complications, like genotyping and demography (see forward Sections 7 & 0), can be incorporated as changes to $P_c(x, a|v^*, s)$.

One such complication is that SNP chips include mostly common variants and all other variants are imputed[1]. Since imputation quality decays rapidly with frequency[2] rare variants are usually omitted from GWAS resulting in an effective (and often literal) MAF cutoff[3]. We model this effect by only including loci with MAF > 0.1%, this changes $n_c(v^*, s)$ to

$$n_c(v^*, s) = \int_{x,a|v>v^*,x>0.1\%} p_c(x, a|s) \quad \text{(A3.12)}$$

and the likelihood retains its form (Equation A3.10).

## 2. Log scale for the distribution of selection coefficients

We choose to parametrize g in terms of $\log_{10}$ s instead of s. Since GWAS restricts the range of selection coefficients we expect to see, g will have a specific scale of (moderate) selection coefficients. This means, that at least on log scale, we expect g to be close to unimodal and extremely well-behaved and we therefore choose to parametrize g in terms of $\log_{10}$ s.

For simplicity, we keep the above notation but henceforth s represents the base 10 logarithm of the selection coefficients. e.g., $s = -2$ takes the meaning of a selection coefficient of $10^{-2}$. Similarly, we use a base 10 log scale for the scaling coefficient c. The transformation back to linear scale is trivial.

## 3. Approximating integrals by Reimann sums

The integral over s in log likelihood (Equation A3.11) is unpractical to use for maximization so we have to approximate it. We approximate it as a sum over a dense grid of selection coefficients. Throughout, we replace integrals over s with sums over a dense grid of selection coefficients, that is by Reimann sums. Specifically, we use the transformation

$$\int_s \quad \rightarrow \sum_{\{s_j\}}$$

with the integration set $S_{int} = \{s\}_j = \{-6, -6 + \frac{1}{16}, -6 + \frac{2}{16}, \dots, -1\}$.

## 4. Parametrization of the distribution of selection coefficient

The log-likelihood we wrote is a functional of the distribution of selection coefficients at GWAS hits, g(s). Numerically maximizing the likelihood requires parametrizing g(s) by a finite number of parameters. The most straightforward approach it to discretize the distribution of selection coefficients. That is, choose a set of selection coefficients $\{s_j\}$ and approximate g(s) as

$$g(s) = g_j \delta(s - s_j) \qquad \text{(A3.13)}$$

with δ being the Dirac delta function and $\sum_j g_j = 1$. The log-likelihood then becomes

$$LL\big(\{g_j\}, c|\{x, a\}\big) = \sum_i \log\big(\sum_k P_c(x_i, a_i|s_k) g_j\big). \qquad \text{(A3.14)}$$

For a constant c, we can define a matrix $P_{i,k} = P_c\big(x_i, a_i\big|s_j\big)$ and rewrite the log-likelihood as a function only of $\{g_k\}$ as

$$LL\big(\{g_j\}\big) = \sum_j \log\big(\sum_k P_{i,j} g_j\big). \qquad \text{(A3.15)}$$

The log-likelihood will take this form for every parameterization in which g(s) is approximated by a finite linear combination of functions of s.

This class of models are called generalized linear mixed models and represent situations in which a variable can arise from one of several distinct distributions. $\{g_j\}$ is known as the mixing distribution and inferring it from data is far from trivial. It has been shown, that because the log-likelihood is a sum of concave functions of linear sums of the parameters of interest the maximum likelihood estimator of $\{g_j\}$ has limited support, meaning only a subset of $\{g_j\}$ would usually be assigned non-zero weights[4].

Indeed, when we try to simulate GWAS models with discrete selection coefficients and infer the distribution of selection coefficients by maximizing the likelihood (Equation A3.15), we see that for virtually all parameter ranges the inferred distributions of selection coefficients are highly irregular, with most selection coefficients given a weight of zero (Fig. A3.1).

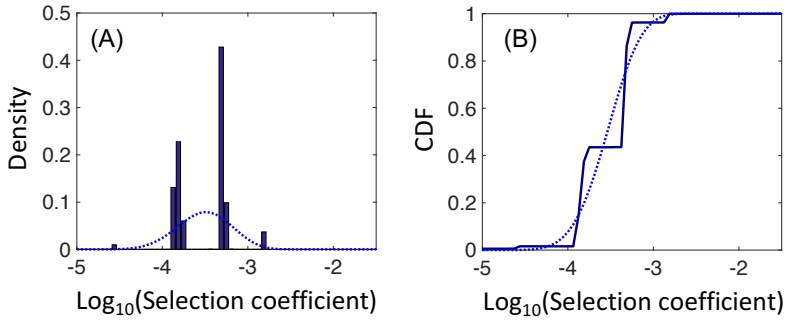



**Figure A3.1.** Example of the irregular distributions inferred for discretized parametrization of g(s), the PDF shown in (A) and the CDF in (B). Shown for a single gamma with $\theta = 0.003$ and $k = 1$. $v^* = 10^{-4}$, $c = 0.5$ and $N_e = 10{,}000$.

## 5. Log Spline Method

A common approach to deal with such problems in generalized linear mixed models is by log spline parametrization[5]. In this approach, the log of a continuous mixing distribution is

parametrized as a spline. The location and number of the spline knots control the (local) smoothing of the mixing density and the parameters of the spline are inferred from the data via maximum likelihood.

We therefore parametrize the distribution of selection coefficients for genome-wide significant variants as

$$g(s) = \frac{\exp[-B(s)]}{\int_s \exp[-B(s)]} \quad \text{(A3.16)}$$

with $B(s)$ being a cubic spline. That is, there are K knots $\{s_k\}$ such that between knots $B(s)$ is a cubic function of s. $B(s)$ and its two first derivatives are continuous across knots and beyond the specified knots $B(s)$ is linear.

The function $B(s)$ is fully specified by its values at the knot location, i.e. by the sequence $\{B_k\}$ s.t. $B_k = B(s_k)$. We can therefore write $B(s)$ for every s as a function of $\{s_k\}$, the knot locations, and $\{B_k\}$, the function values at those knots, - $B(s|\{s_k\},\{B_k\})$. This allows us to write the likelihood in terms of the finite set of parameters $\{s_k\}$ and $\{B_k\}$:

$$LL(\{s_k\},\{B_k\},c|\{x,a\}) = \sum_i \log\left(\int_s P_c(x_i,a_i|s)g(s)\right) = \sum_i \log\left(\int_s P_c(x_i,a_i|s)\frac{\exp[-B(s|\{s_k\},\{B_k\})]}{\int_s \exp[-B(s|\{s_k\},\{B_k\})]}\right). \quad \text{(A3.17)}$$

Any analytic function can be approximated by a large enough number of knots, K. However, since we expect $g(s)$ to be smooth and well behaving we should not need a large K. Therefore, we introduce a penalty on K into the likelihood

$$LL_K(\{s_k\},\{B_k\},c|\{x,a\}) = LL(\{s_k\},\{B_k\},c|\{x,a\}) - \frac{1}{2}\log(n)\,K, \quad \text{(A3.18)}$$

similar to a BIC penalty, with n being the number of data points.

Note, that the likelihood should only very weakly depend on the exact location of the knots, $\{s_k\}$. The reason for this is that when the knots are sufficiently dense the $g(s)$ changes slowly between knots and many different knot locations could capture this change. That is why we do not consider knot locations as extra parameters in our penalized likelihood, i.e. we use K and not 2K in Equation A3.18.

## 6. Maximizing the likelihood

We have discovered that maximization is extremely efficient when c and $\{s_k\}$ are held constant, i.e. when $\{B_k\}$ are the only parameters. The reason that holding c as a constant greatly accelerates the maximization is because it allows us to pre-calculate $P(x, a|s)$ at each data point for each selection coefficient in the integration set, $S_{int}$. Holding the knot locations, $\{s_k\}$, as constants greatly speeds up maximization because the likelihood only weakly depends on knot locations. Therefore, maximization is very slow over the highly-degenerate parameter space created by allowing knot locations to change.

We therefore use a two-stage maximization scheme: we maximize the (penalized) log likelihood separately for a set of c and $\{s_k\}$ and we choose sets of c and $\{s_k\}$ using an MCMC sampling method. This process is very similar to simulated annealing optimization of AIC (SALSA) as described by Hansen & Kooperberg[6].

In the first stage, we only maximize over $\{B_k\}$. We pre-calculate $P_{i,j} = P(x_i, a_i|s_j)$ for every data point $(x_i, a_i)$ and for every $s_j \in S_{int}$. Every function evaluation involves spline interpolation of $B(s)$ from $\{B_k\}$ to $\{B(s_j)\}$ for every $s_j \in S_{int}$. The log-likelihood is then

$$LL(\{B_k\}|c, \{s_k\}) = \sum_i \log\left(\frac{1}{N}\sum_j P_{i,j} \exp(-B(s_j))\right) \quad \text{(A3.19)}$$

with $N = \sum_j \exp(-B(s_j))$. We find the values of $\{B_k\}$ that maximize the log likelihood via the Nelder-Mead algorithm[7]. We mark this maximum likelihood as $LL^{max}(c, \{s_k\})$.

In the next stage, we sample $c$ and $\{s_k\}$, using a Metropolis-Hastings algorithm[8], from a distribution proportional to their marginal penalized likelihood

$$\Pr(c, \{s_k\}) \propto \exp\left(LL^{max}(c, \{s_k\}) - \frac{1}{2}\log(n)\, K\right)$$

$$= \exp\left(LL^{max}(c, \{s_k\})\right)\left(\frac{1}{\sqrt{n}}\right)^K \qquad \text{(A3.20)}$$

with n again being the number of data points.

We use the following procedure to propose new values, $c^*$ and $\{s_k^*\}$. At each iteration, we make one of these four proposals:

1. Propose a new $c^*$ by drawing a normal variant $\nu \sim N(0,1)$ and setting $c^* = c - 0.1\,\nu$ (note, that c can be negative because it is on a log scale).
2. Move one of the knot $s_j$ to a new position $s_j^*$ (see details below). All other knots remain the same.
3. Remove one knot, chosen uniformly among existing knots.
4. Add a knot (see details below).

For most iterations we choose proposals 1,2,3 and 4 with probabilities

$$\left[0.1, 0.45, 0.45\frac{\sqrt{n}}{\sqrt{n}+1}\, 0.45\frac{1}{\sqrt{n}+1}\right] \qquad \text{(A3.21)}$$

respectively. However, to speed up the initial burn in period we make sure that if a new value for c is proposed an accepted in one iteration the next iteration would also propose a new value for c, i.e. use proposal 1.

When a new knot is added it is chosen uniformly in the range [-6,-1], excluding positions near existing knots. This exclusion is necessary to avoid numerical instability resulting from highly fluctuating behavior when two knots are very close. We chose the minimal distance between knots to be 0.5 by experimenting on simulated data. Moving a knot consists of removing a knot at random and adding a new one according to the above procedure.

We accept $c^*$ and $\{s_k^*\}$ with probability

$$\min\left[1, \exp\left(LL^{max}(c^*, \{s_k^*\}) - LL^{max}(c, \{s_k\})\right)\right] \qquad \text{(A3.22)}$$

following the Metropolis-Hastings algorithm[8]. Note, that since in the above procedure the probability to remove a knot is $\sqrt{n}$ times larger than the probability to add a knot (Equation A3.21), the penalty term in Equation A3.20 is canceled out in Equation A3.22. This increased probability to remove a knot not only simplifies Equation A3.22 but also greatly improves computational efficiency by avoiding configurations with larger numbers of knots for which the optimization is much slower.

We repeat this procedure for 2,000 iterations and choose the values of c, $\{s_k\}$, $\{B_k\}$ with the maximal penalized log-likelihood $LL_K(\{s_k\}, \{B_k\}, c|\{x, a\})$. These values can be translated to the desired maximum likelihood estimates of $g(s)$ via Equation A3.16.

## 7. Non-equilibrium demography

We use simulations based on a haplotype-derived demographic model to incorporate the effects of non-equilibrium demography into our inference. As detailed in Appendix 2, we pieced

together the historic European population size as inferred from two and four European haplotypes[9]. We use our forward simulator to simulate this model, see details in Appendix 2.

We use the simulations to estimate our likelihood under this model. We run the simulations 240,000,000 times (simulating 240Mbp) for each selection coefficient in $S_{int}$. This allows us to get a detailed SFS for each selection coefficient. We bin the MAF into 40 bins with bin edges

$$\{10^{-3}, 10^{-2.98}, 10^{-2.96}, 10^{-2.94}, 10^{-2.92},$$

$$10^{-2.9}, 10^{-2.8}, 10^{-2.7}, 10^{-2.6} \dots, 10^{-1.1},$$

$$0.1, 0.125, 0.15, 0.175 \dots, 0.5\}.$$

The very small bin sizes close to $x^*$ prevent numeric problems at strong selection, when all detected SNPs are very close to $x^*$. At low frequencies, bin edges are on a log scale so at nearly neutral sites each bin will have a similar number of variants. At high frequencies, we use a linear scale to fully capture the bottleneck's effect on the frequency distribution.

We can then replace each MAF in our data, $x_i$, with its corresponding bin $b(x_i)$, where we mark by $b(x)$ the function that returns the bin number the MAF x belongs to. Then we can replace $\tau(x_i|s)$ in our likelihood with the probability that a simulation ends with a segregating allele at bin $b(x_i)$, $P(b(x_i)|s)$, as estimated by our simulations.

We can then calculate the expected number of discovered hits per site for each selection coefficient, $n_c(v^*, s)$. The probability of discovery for a variant at frequency x and with selection coefficient s is

$$1 - \mathrm{erf}\left(\sqrt{\frac{1}{2cs}\frac{v^*}{2x(1-x)}}\right). \qquad \text{(A3.23)}$$

We weigh each frequency in our full simulated SFS by this factor and sum over it to produce $n_c(v^*, s)$. Then $P_c(x_i, a_i|s)$ in our log-likelihood (Equation A3.11) takes the form

$$P_c(x_i, a_i|s) = \frac{P(b(x_i)|s)\, p_c(a|s)}{n_c(v^*,s)}. \quad \text{(A3.24)}$$

## 8. Simulating datasets

In order to evaluate the accuracy of our inference method, we run it on simulated datasets under known parameters. We aim to test the inference on a wide range of selection coefficient distributions. For each simulated variant, we draw a selection coefficient from a combination of gamma distributions. This allows us to simulate both simple, unimodal distributions and complex and/or multimodal distributions. We then draw the variant's frequency and effect size, and retain the variant if it contributes more than $v^*$ to the variance and its MAF is above $x^*$. Note, that the resulting distribution of selection coefficients at discovered variants, which we are attempting to infer, will no longer be a combination of gamma distributions. We also tried a variety of different values for the constant $c$.

For a constant population size, we can draw frequencies directly from our analytic results while for non-equilibrium demography, we have to rely on simulations. For a constant population size, we draw segregating allele frequencies with probability proportional to $\tau(x|s)$, conditional on $x > x^* = 0.1\%$. Therefore, the frequency distribution is

$$P(x|s) = \frac{\tau(x|s)}{\int_{x>x^*} \tau(x|s)} \quad \text{(A3.25)}$$

with $x > x^*$, and we draw from it using the acceptance-rejection method. For non-equilibrium demography, we draw the frequencies from our simulated SFS (see Section 7 above) with $x >$

$x^*$. Once we have the frequency, we can draw an effect size from Equation A3.4 while conditioning on a $> \sqrt{v^*/2x(1-x)}$ to make sure that $v > v^*$.

For non-equilibrium demography, we only simulate the SFS for selection coefficients within the integration set $S_{int}$. Therefore, those are the only selection coefficient we can draw frequencies from. We therefore, round each drawn selection coefficient to the closest selection coefficient within $S_{int}$.

We draw 2,000 variants for each dataset. We use $N_e$=10,000 as our constant population size. We use $v^* = 10^{-4}$, to keep with the order of magnitude of real GWAS threshold variances, and $x^* = 0.1\%$, to reflect the quality of imputation currently achieved in large GWAS. We use $c = 0.1666, 0.25, 0.5, 1$ or $2.5$, with a constant population size these values correspond to corresponding to $v^* = 3v_s, 2v_s, v_s, 0.5v_s$ and $0.2v_s$ respectively, with $v_s = 4c/2N_e$ being the expected contribution of a strongly selected allele (see Chapter 2). That is, for small c the threshold is very large and only variants with extreme contributions to the variance would be captured, for large c most of the variance from moderately selected sites is captured and for moderate c the variance is partially captured, as in real GWAS.

For each value of c, we draw selection coefficients from one of five distributions. Either from an exponential distribution (gamma distribution with shape factor 1) with mean selection coefficient of 1e-4, 3e-4 or 1e-3. Or from a bimodal distribution created from combining 3 gamma distributions (one with mean 1e-4 and shape 5, one with mean 3e-4 and shape 0.3 and the third with mean 1e-3 and shape 5). Or from a unimodal combination of 3 gamma distribution (one with mean 1e-4 and shape 5, one with mean 3e-4 and shape 1 and the third with mean 1e-3 and shape 0.3)

## 9. Predictions for GWAS of Increased Study Size

We can only make predictions from the range of selection coefficients observed in GWAS. We define this range as the range which contains all but the 100 most extreme selection coefficients. That is, the range at which the CDF of the inferred $g(s)$, $G(s) = \int_{s'<s} g(s')$, is between $50/n$ and $1 - 50/n$. We define $s_{down}$ and $s_{up}$ as the lower and upper bound of this range, that is $G(s_{down}) = 50/n$ and $G(s_{up}) = 1 - 50/n$. This choice is designed to use all the selection coefficients for which we have information.

We can predict the number of variants to be discovered from this range as study sizes increase, that is as $v^*$ decreases. To do this, we first need to know the density of sites at this range, which is $\rho(s) = n\, g(s)/n_c(v^*, s)$, with n bring the number of GWAS hits. Then the expected number of variants discovered as GWAS increase is simply

$$n(v) = \int_{s_{down}}^{s_{up}} \rho(s)\, n_c(v, s)$$

with v bring the new threshold variance.

We can define $v_c(v^*, s)$ as the amount of variance explained by a newly arising mutation with selection coefficients s in a GWAS with threshold variance $v^*$. Therefore,

$$v_c(v^*, s) = \int_{v>v^*, x>x^*} 2a^2x(1-x) \cdot P_c(x, a|s)$$

and the expected amount of variance explained by an expanded GWAS from selection coefficients between $s_{down}$ and $s_{up}$ is

$$V(v) = \int_{s_{down}}^{s_{up}} \rho(s)\, v_c(v, s).$$

Though these are only expectations, the variance around these expectations is negligible because of the large number of variants involved.

We compare these results to simulations in which we extend our previously simulated datasets down to $v^* = 10^{-5}$ to simulate a 10-fold increase in study size.

## 10. Future Application to data

Applying our methods beyond simulated data to results from actual GWAS would induce further considerations. GWAS do not produce a simple list of causal, completely independent variants. Instead, GWAS produces an estimated effect size for each included SNP, and those SNPs can be at very strong LD with each other.

The simplest issue to deal with is error in effect size estimation. This error is, to a good approximation, normal[3] and the estimated effect size, $\hat{a}$, is distributed as

$$\hat{a} \sim N\left(a, \sqrt{\frac{1}{m} \cdot \frac{1}{2x(1-x)}}\right). \qquad \text{(A3.26)}$$

with m being the (effective) study size. This error can therefore be explicitly taken into account in the likelihood, as it only results in a small change to $P_c(x, a|s)$. Similarly, we could take into account sampling errors in allele frequency estimation, except that with study sizes in the hundreds of thousands these errors are insignificant.

A much more problematic issue is identifying good proxies for causal variants[10]. Since each causal, trait-affecting variant has many nearby variants in LD with it, GWAS capture multiple adjacent significant hits for each causal locus. Though these hits are all at similar frequencies, the process by which we choose the tagging variant increases our uncertainty in our estimation of the causal allele frequency and effect size and may also bias their distribution. For example,

choosing the most significant SNP in a region may bias upwards the estimated effect size. This process also has the potential of misidentifying the ancestral state of an allele, e.g. if there are two adjacent alleles in LD and the derived allele frequency of one is close to the ancestral allele frequency of the other. This is the reason why we choose to work with minor allele frequency and not derived allele frequency.

The situation becomes worse when there are multiple causal alleles in the same region. In this case, inferring the correct number of true signals is very difficult. For example, two causal alleles in LD with each other may look like one, larger causal signal and vice versa. This problem becomes worse as GWAS power increases and more and more signals are discovered since more and more of them are in close proximity to each other. There have been multiple attempts to mitigate this problem but it still far from resolved[10-13].

The linkage disequilibrium between variants also means that they are not statistically independent data points. Since our model does not directly account for LD at the moment, we use the likelihood derived from our single variant model which now serves as a composite likelihood. Estimating confidence intervals would now require bootstrapping over approximately independent blocks in the genome rather than individual variants[14]. Future theoretical work is needed to address the effects of linkage on allele dynamics in our model.

Another possible factor to consider is the effects of mutation bias or recent directional selection on our estimated distribution of selection coefficients. However, as shown in Chapter 2, these most likely have had small effects on the overall distribution of frequency and effect sizes. This is in contrast to tests of directional selection which sum over many variants the small changes in frequency caused by these phenomena. This is also the reason why, unlike tests of directional

selection, we do not expect this work to be sensitive to small, subtle biases in allele frequencies caused by population structure.

## 11. Additional Figures

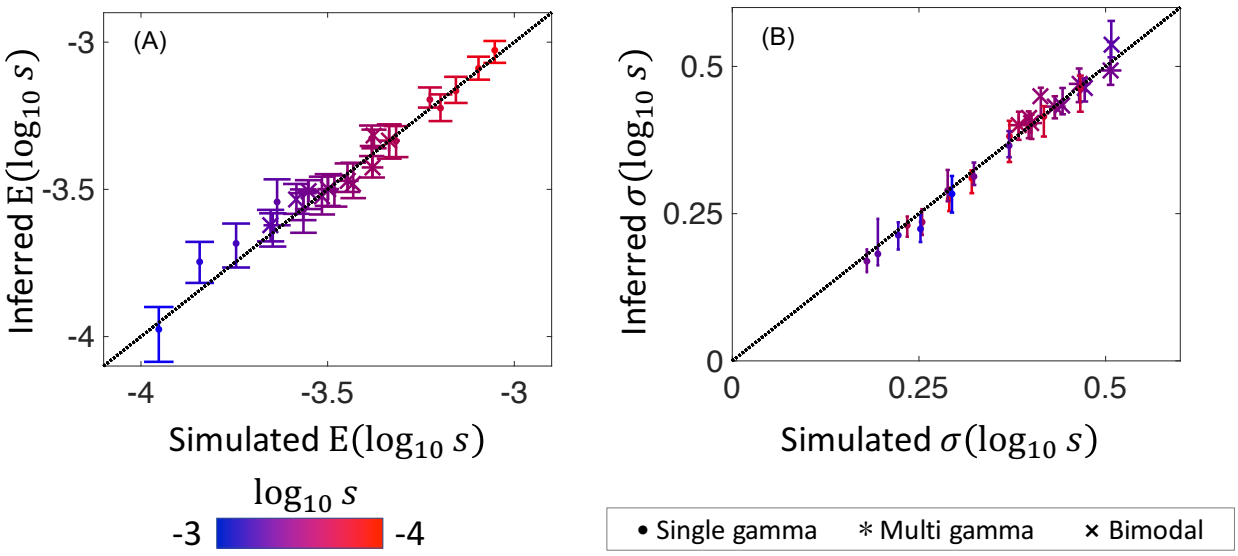



**Figure A3.2.** Inferred mean and standard deviation of the distribution of logged selection coefficients under non-equilibrium demography are good estimators of the simulated mean and standard deviation. (A) Inferred mean log10 of selection coefficients vs. the simulated mean. (B) Inferred standard deviation log10 of selection coefficients vs. the simulated standard deviation. The real mean log10(s) of the simulated datasets are marked by color gradient from blew (log10(s)=-3) to red (log10(s)=-4) and the shape of the distribution by the shape of the marker (a point for a single gamma, an asterisk for multiple gammas and an 'x' for bimodal. Shown for a single simulation for each combination of the five distributions of selection coefficients at newly arising mutations and five values of c. All simulated with $v^* = 10^{-4}$ and our non-equilibrium demographic model. See Figure 3.2 for similar results for equilibrium demography.

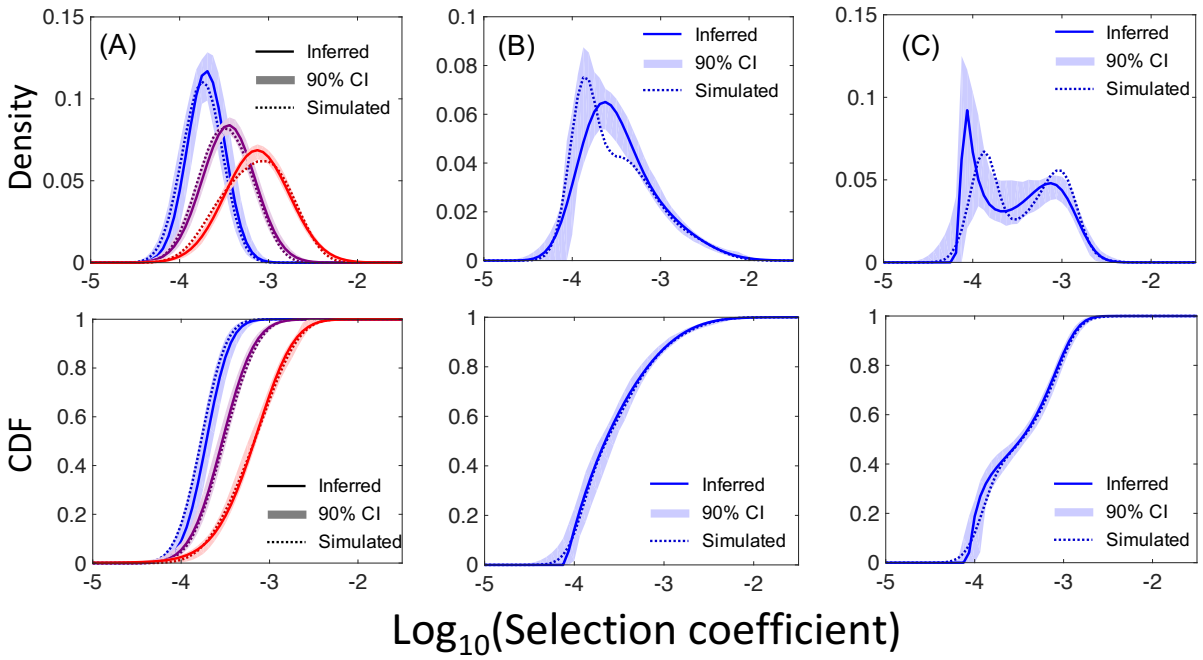



**Figure A3.3.** The distribution of selection coefficients as inferred from simulations under non-equilibrium demography (solid curves + 90% CI in shaded regions) compared to the simulated distribution (dotted curves). (A) Single gamma distributions of selection coefficients for newly arising mutations. (B) Unimodal mixture of gamma distributions of selection coefficients for newly arising mutations. (C) Bimodal mixture of gamma distributions of selection coefficients for newly arising mutations. All with $c = 0.5, v_s = 10^{-4}$ and our non-equilibrium demographic model. See Figure 3.3 for similar results for equilibrium demography.

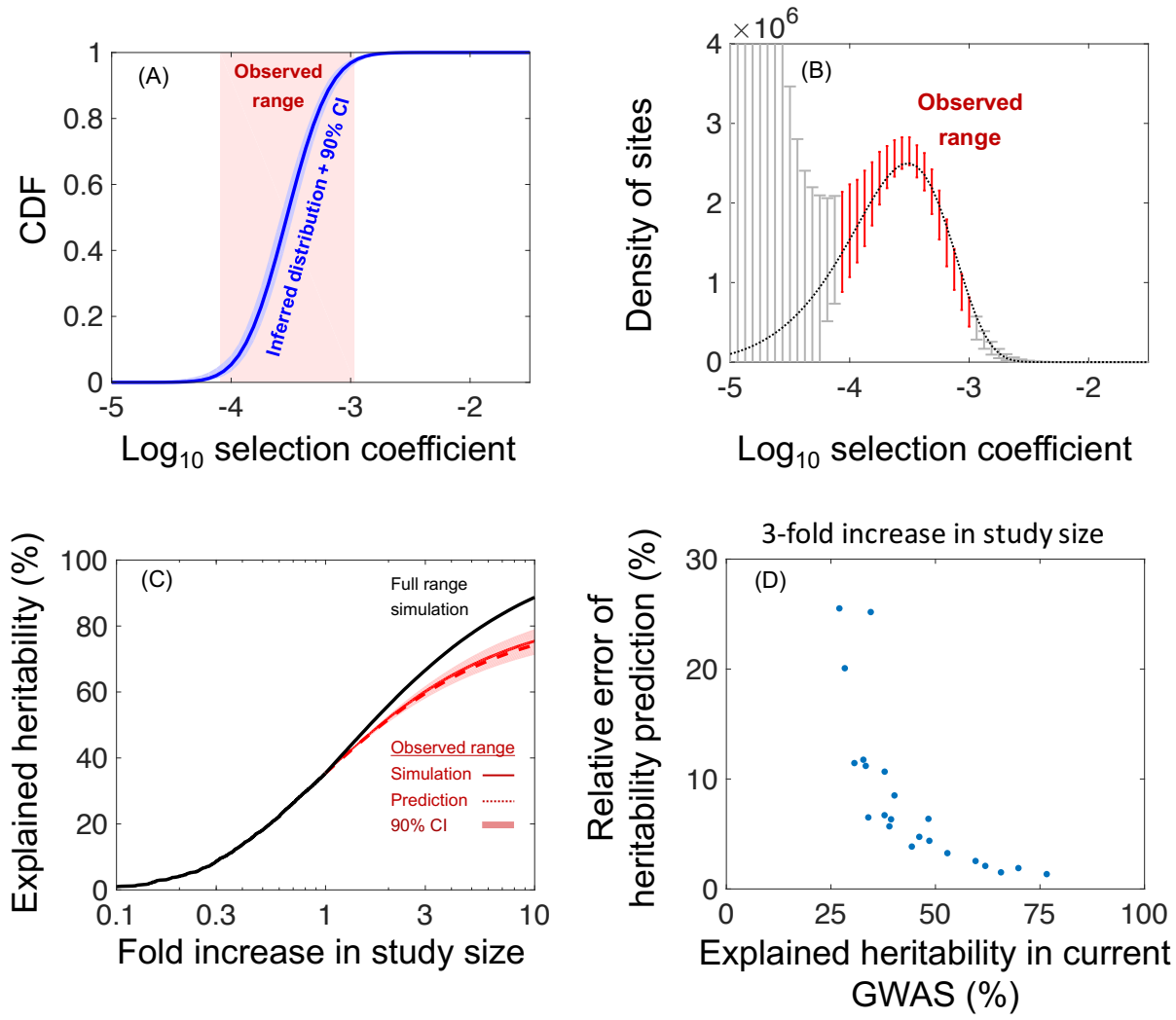



**Figure A3.4.** Predictions based on the inferred distributions under non-equilibrium demography. (A) The inferred CDF of selection coefficients for GWAS hits with the observable range, defined as the range containing all but the 100 most extreme observed selection coefficients. (B) The density of sites as function of selection coefficient, with the observable range marked in red. The density in the observable range is used to make predictions for future GWAS. (C) The increase in explained heritability as a function of the increase in study size. The prediction and contribution from the observed range are in red and simulation results for the full range are in black. (D) Relative error of heritability prediction (ratio of 90% CI to predicted heritability) for a 3-fold increase in study size as a function of explained heritability in current GWAS. (A-C) are for $c = 0.5$ with an exponential distribution of selection coefficients for newly arising mutation with mean $E(s) = 3 \cdot 10^{-4}$. All results are with $v^* = 10^{-4}$ and our non-equilibrium demographic model. See Figure 3.4 for similar results for equilibrium demography.

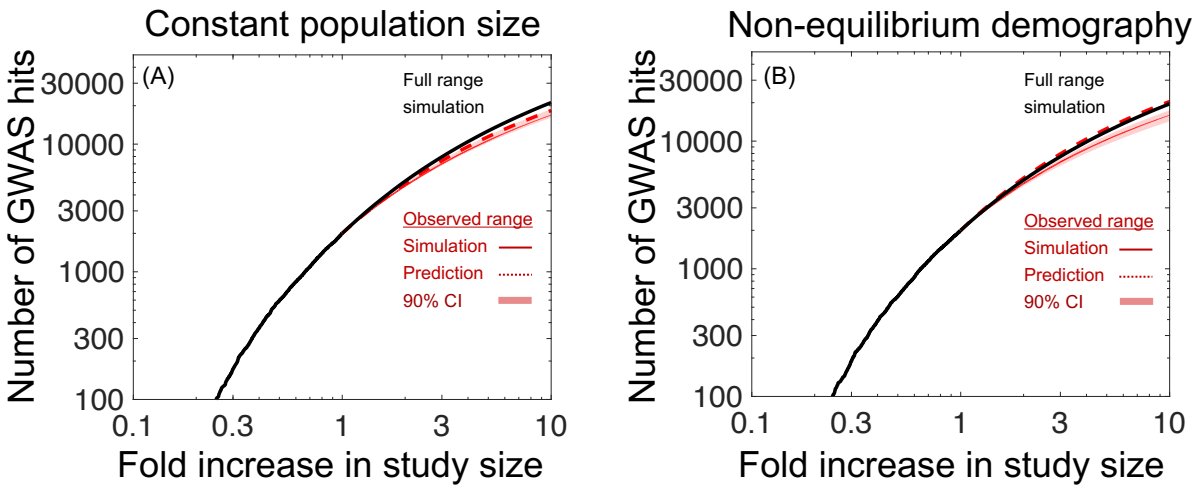



**Figure A3.5.** Predictions of the number of variants identified in GWAS ("GWAS hits") as a function of study size. (A) For a constant population size of $N_e = 10{,}000$. (B) For our non-equilibrium demographic model. The prediction and contribution from the observed range are in red and simulation results for the full range are in black. Results shown for $c = 0.5$, $v^* = 10^{-4}$ with an exponential distribution of selection coefficients for newly arising mutation with mean $E(s) = 3 \cdot 10^{-4}$.

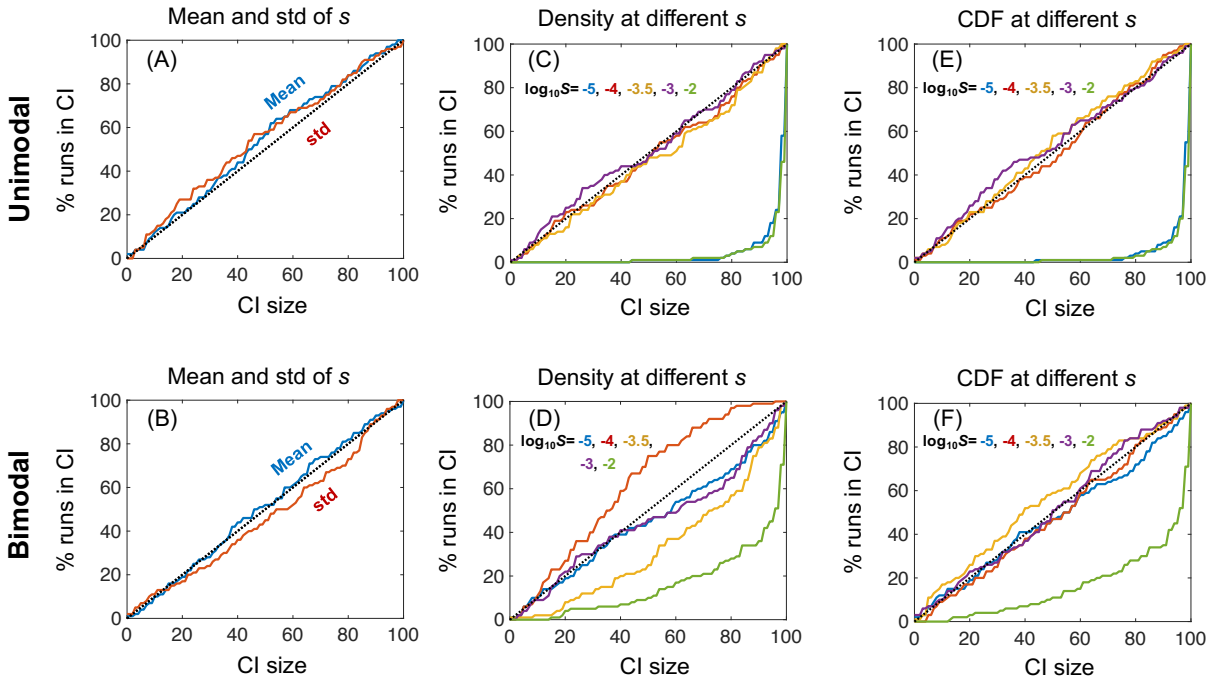



**Figure A3.6.** Calibration of confidence intervals for equilibrium demography. The proportion of runs within CI should be equal to CI size for well-calibrated CI. (A-B) Calibration of CI for mean and standard deviation of the distribution g(s). (C-D) Calibration of CI for the density g(s) at different values of s. (E-F) Calibration of CI for the CDF of g(s) at differnet values of s. (A,C,E) are for a single gamma distribution with $E(s) = 3 \cdot 10^{-4}$ while (B,D,F) are for our bimodal distribution. Results shown for $c = 0.5, v_s = 10^{-4}$ and a constant population size of $N_e = 10^4$.

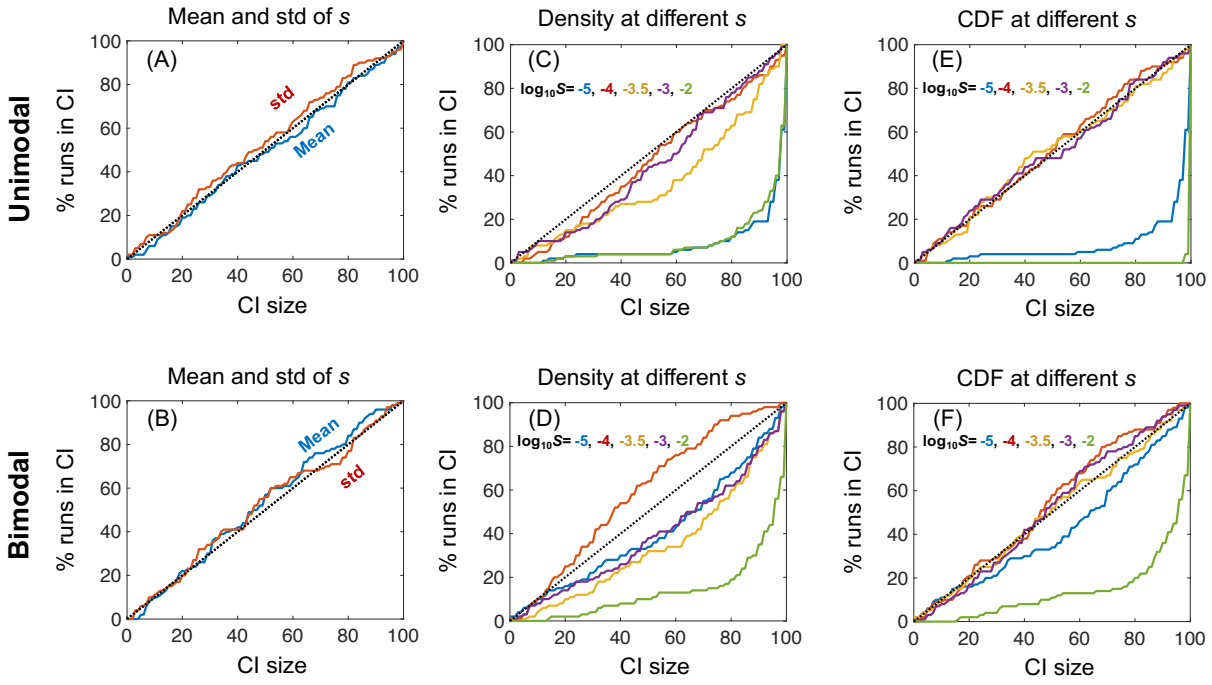



**Figure A3.7.** Calibration of confidence intervals for non-equilibrium demography. The proportion of runs within CI should be equal to CI size for well-calibrated CI. (A-B) Calibration of CI for mean and standard deviation of the distribution g(s). (C-D) Calibration of CI for the density g(s) at different values of s. (E-F) Calibration of CI for the CDF of g(s) at differnet values of s. (A,C,E) are for a single gamma distribution with $E(s) = 3 \cdot 10^{-4}$ while (B,D,F) are for our bimodal distribution. Results shown for $c = 0.5, v_s = 10^{-4}$ and our non-equilibrium demographic model.

## References


1. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).

2. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010).

3. Sham, P.C. & Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* **15**, 335-46 (2014).

4. Lindsay, B.G. The Geometry of Mixture Likelihoods: A General Theory. *The Annals of Statistics* **11**, 86-94 (1983).

5. Kooperberg, C. & Stone, C.J. A study of logspline density estimation. *Computational Statistics & Data Analysis* **12**, 327-347 (1991).

6. Hansen, M.H. & Kooperberg, C. Spline Adaptation in Extended Linear Models (with comments and a rejoinder by the authors. *Statist. Sci.* **17**, 2-51 (2002).

7. Nelder, J.A. & Mead, R. A simplex method for function minimization. *The computer journal* **7**, 308–313 (1965).

8. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109 (1970).

9. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925 (2014).

10. Spain, S.L. & Barrett, J.C. Strategies for fine-mapping complex traits. *Human Molecular Genetics* **24**, R111-R119 (2015).

11. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).

12. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* (2014).

13. Zheng, J. *et al.* HAPRAP: a haplotype-based iterative method for statistical fine mapping using GWAS summary statistics. *Bioinformatics* **33**, 79-86 (2017).

14. Berisa, T. & Pickrell, J.K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283-285 (2016).